Aalto University
School of Engineering

Petr Hajduk

Big Data for Activity Based Transport Models

Thesis submitted for the examination for the degree of Master of Science in Technology

Espoo 21.11.2018
Supervisor: Assistant Professor Claudio Roncoli
Advisor: Dr. Mikko Pihlatie

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Petr Hajduk	
Title of thesis Big Data for Activity Based Transport Models	
Master programme Spatial Planning and Transportation Engineering	**Code** ENG26
Thesis supervisor Assistant Professor Claudio Roncoli	
Thesis advisor(s) Dr. Mikko Pihlatie	
Date 21.11.2018 **Number of pages** 68 +12	**Language** English

Abstract

Our civilization needs to know as much information about itself as possible in order to keep running. One of the important fields is the field of transportation and since we could not measure all the movements happening on planet Earth, we need transport modelling. As of 2018, for the area of a metropolis the four-step model still seems to be a state of practice of modelling transportation. This comes with several disadvantages such as lack of detail (aggregation to zones) or oversimplifying of the travel demand phenomena (trips are not combined into daily schedules). To remedy these disadvantages, the scientific community came up with activity-based models that addressed those issues. The increased level of detail has however increased the demand for data. Nowadays the data is obtained from costly travel surveys that make the methodology less viable option for the practitioners. Therefore, in this thesis the focus are possible new sources of data for the model and using the open datasets to build an activity-based model.

First, we examine the existing big data sources and evaluate their usefulness for the model. As a result of this evaluation, we carry on to create synthetic data handling the movements of the studied population, as no big data source related to movement of people was found useful for creating the data suitable for the model.

We used the Capital region of Helsinki, Finland as a region for the case study to deal with the real data environment. The data is generated by disaggregation of statistical data aiming at preserving the variability in a maximum achievable way. Where needed, assumptions are used to forward the process.

Using the synthetic big data a transport model was created. Despite the fact that the accuracy of the model in terms of error on link volumes does not reach the level of some other previously developed models, it is still surprisingly precise regarding the idea that solely open data and statistics were used. In the discussion possible synergies with other big datasets is described with respect to the experiences from the case study.

Keywords activity based model, transport engineering, big data, mobile data

Acknowledgements

Many thanks to Assistant Professor Claudio Roncoli, who helped me to get through the thesis as well as VTT that funded the thesis. Without their support my stay in Finland would be at stake. VTT also supported me with their environment and their devices that I used during my thesis. From Aalto I would also like to thank Milos Mladenovic for the inspiration for this topic as well Mikko Pihlatie from VTT to have the courage to support me. I also need to thank my family for enabling my stay in Finland despite several hardships.

Petr Hajduk

Espoo 21.11.2018

Table of Contents

Abstract
Acknowledgements
Table of Contents i
List of Abbreviations iii
1 Introduction 1
2 Travel Demand Modelling Approaches 3
2.1 The Idea of Transport Modelling 3
2.2 Categorization 3
2.3 Scales of Models 4
2.3.1 Microscopic 4
2.3.2 Macroscopic 4
2.3.3 Mesoscopic 4
2.4 Four Step Modelling 5
2.5 The Critique of the Four Step Models 5
2.6 Activity Based Modelling 5
2.7 Implementing Activity Based Modelling in Matsim 6
2.8 Comparing the Resources Needed for Both Methodologies 7
2.9 Data Requirements 8
3 Big Data 8
3.1 Possible Big Data Sources 9
3.1.1 Mobile Networks - Call Detail Records 9
3.1.2 Use Cases of CDR for Transportation Engineering 10
3.1.3 Social Networks 11
3.1.4 Openstreetmaps 12
3.1.5 Route planners 13
3.1.6 Google Maps 13
3.1.7 Mobile phone applications 14
3.1.8 GTFS 15
3.2 Comparison of the Existing Big Data Sources 15
3.2.1 Evaluation Criteria 15
3.2.2 Evaluated Datasets 16
3.3 Conclusion for Big Data Sources 17
4 Traditional Data Sources 18
4.1 Travel Diaries 18
4.2 Transport Surveys 18
4.3 Population Census 19
4.4 General Statistics about Population 19
4.5 GIS Data 20
4.6 Note on Elevation Data 20
5 Test case – Helsinki, Finland 21
5.1 Basic needs of the Capital region model 21
5.2 Loading the network 22
5.2.1 Basic Topology 22
5.2.2 Format and Coordinate Systems 22
5.2.3 Adjustments for Public Transportation and Cycling 22
5.2.4 Generating the network 23

5.3 Loading the Public Transportation 23
5.3.1 Public Transportation Implementation in Matsim 23
5.3.2 Generating Public Transport Files 24
5.4 Generating the Population 24
5.4.1 What Do We Need to Create? 24
5.4.2 Mobile Data as a Shortcut 25
5.4.3 Statistics 25
5.4.4 Synthetic Big Data 25
5.5 Implementing the Model 26
5.5.1 Generating the Network with Public Transport 26
5.5.2 Generating the Population 27
5.5.3 Testing of the Generated Population 32
5.5.4 General Commentary on Testing 38
5.5.5 Generating the Plans 38
5.5.6 Joining the Data Together 48
5.5.7 Running the model 52
6 Results and Discussion 56
6.1 Scanning the Big Data Environment 56
6.2 Modelling the Capital Region 56
6.3 Possible Improvements 59
6.4 Possible Use Cases for the Developed Model 59
6.5 The Take-Away Message for the Big Data Environment 60
7 Conclusion 62
References 64
Appendices I
Appendix 1. Excerpt of twitter data II
Appendix 2. Code extracting tweets with coordinates and time, plotting them afterwards III
Appendix 3. Households in Capital Region by Size and Statistical Unit IV
Appendix 4 Population and Plans Generation Code Scheme XI
Appendix 5 The distribution of work activity starts vs ends XII

List of Abbreviations

AADT	Annual average daily traffic
CDR	Call detail record
FSM	Four step model
GTFS	General Transit Feed Specification
ID	Identification number
MNO	Mobile network operator
OD matrix	Origin-destination matrix
OSM	Openstreetmap.org
Reittiopas	Reittiopas journey planner
Statfin	Statistics Finland
TAZ	Transport analysis zone
VTT	Teknologian tutkimuskeskus VTT Oy (VTT Technical Research Centre of Finland Ltd)

1 Introduction

Initially, people had little need to move around and the only activity that would take place outside of their (temporary) homes were hunting and gathering. As the population developed, the activity framework grew in variety, people needed to go to the market, have their hammer fixed at the blacksmiths' and visit their relatives in the neighboring town, which generated travel demand. As the first roads started to appear, so did the congestion, when the demand exceeded the offered road capacity. The evidence for first traffic problems and their solutions can be traced back to the Ancient Rome (Van Tilburg, 2011). With the increased mobility thanks to the invention of the new transport modes during the industrial revolution, cities were able to increase their size, which in turn put the daily activities even further away from each other.

Due to the mere size of the transportation network in the expanded cities it became harder and harder to deal with the increasing traffic with simple methods. As a remedy, the traffic (vehicle flow) was first measured creating the traffic model and finally the transport demand was modelled into a transport demand model. The first transport demand models were developed for Chicago and reflected mostly car network. The methodology for the model became later known as the four-step model, the scientific branch dealing with these models became known as Transport Modelling or Traffic Forecasting. The aim from thereafter was to increase the precision of traffic flow predictions and the efficiency of modelling. A resulting model is always a trade-off between these two.

Two methodologies are commonly used in the field of Transport Modelling. The four-step model being developed earlier and an activity-based model developed later. The four-step model is continuous in terms of dealing with traffic flows, while activity-based is discrete since it is based on agents. A four-step model is arguably easier to deploy as it does not demand too much data inputs while activity-based model needs very detailed input. (Zhong, Shan, Du, & Lu, 2015)

Due to the extra costs connected with making models that are more granular, only very general four-step models are usually being used at the level of city traffic. These are so called macroscopic models. Such models use traffic analysis zones or in exceptional cases facilities as a basic spatial unit and the population of the zone as a unit of population. Single trips are not connected into tours and daily travel patterns, capacities of facilities such as restaurants are often neglected (if travelling for leisure is even taken into account) and classification of the trips is usually relying on (non-)home or (non-)work based trips. This leaves out lots of complexity and makes the model oversimplified for many situations. Furthermore, car ownership is only taken into account at the level of zones, but not for a single individual or a household. (Mladenović MN, 2014)

Time-scale is another issue with such models as usually only rush-hour traffic is modeled; neglecting that some links might get overloaded in a very different time of the day (midnight, weekends etc.) The departure time is only added after the first two steps in the sense of a filter. (McNally, 2000) Such oversimplified model is then presented to the authorities to support their decisions. This implies that the cities are mainly designed to serve the morning commute.

Another problem with the FSM is that they can hardly show the reason for congestion. Even if one breaks down the model to zone-to-zone flows, it is impossible to tell exactly

why the people desire to use such a link. This makes it impossible for city planners to use such measures as mixed zones, local services and increased walkability to reduce the congestion simply because the zone includes everything from a restaurant to a factory. In addition to that, in Helsinki we might observe that the afternoon congestion is the most severe when the weather is good. This would imply that people simply stay more at home when it rains. I do not have any exact measurements for this claim though, but my intention was to show that the cause of the congestion might be more trivial than we expect. This level of detail is however impossible to model with the current four-step model. For more comparison between the methodologies and their outcomes, see (Zhong, Shan, Du, & Lu, 2015).

To address these problems, mesoscopic activity-based models have been proposed since long ago; see (Kitamura, 1988). These models use a different philosophy as they aim to model the individual and its behavior. This tends to be rather complex as one must develop the daily routine for each individual and then combine the data with the information about the network. This process is quite cumbersome as there is no data available at the level of each individual and as we figure out later, it is not even desirable to model each individual 100 % precisely due to privacy issues and overfitting. Despite all these hurdles, the model can answer most of the issues criticized above and can give you much more. Once one figures out the movements of people with the desired precision, not only you know the annual average daily traffic (AADT) for each link and how many people will be probably using a certain bus line, you get to know the reasons for people's mobility, their mobility patterns, population allocation through time, and cause of the congestion (with precision to age distribution, activity people are pursuing or the role of the link within the tours of the individuals).

In order to model the population and its daily plans with desired variation, we need to create a dataset that can be considered big data – all the people to be simulated with their travel patterns and initial mode of transport, check for Matsim requirements in section 5.4. My idea was to include existing big data in the process, as seen in section 3.3. The tempting feature of big data is definitely the variation and granularity it has although this is also its major setback as it contradicts any privacy (Porras, 2006, p. 61). The big data regarding humanity can also never be exact reflection of the reality as it will always involve error due to lower than full penetration of population or in the case of mobile data people might simply switch off their phone for a while and be therefore missing from the data. It seems impossible that everybody would be forced to make their phone connected to the network all the time, at least not as of now, in year 2018.

Another source of information is statistics. Statistics is aggregated data and getting the raw data usually comes with the issues of privacy, overfitting and upscaling bias. Imagine for example upscaling the location of leisure facilities, or any other phenomena that are rare enough to barely appear in the statistical sample.

However, overcoming the full picture should not be that difficult when one imagines the aggregated data as a reflection or a shadow of an object. For the model, we need to recover the object using only those shadows. We do not care that we do not get 100 % picture, but it should be within good precision for its volume, dimensions, weight etc. Therefore, we take the object and use multiple shades to study it and in the end, we should be able to get an object not dissimilar to the original object - this will be the big data that we will load into the model. See Figure 1 for an illustration of the idea.

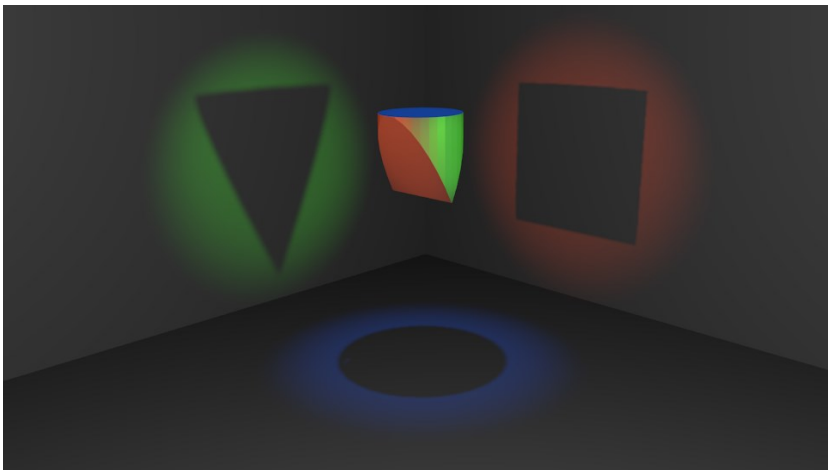


Figure 1 Idea of reconstructing the image based on multiple shadows. Source: https://prostart.me/

In my thesis, I focus on the area of Greater Helsinki (Helsinki, Espoo, Vantaa, and Kauniainen) also called Capital region (Pääkaupunkiseutu in Finnish). Having been living in the area for a second year already this focus is not a random choice as I am quite familiar with the area right now. Furthermore, VTT, that helped me to conduct this thesis, needs a good source of various data such as passenger loads, levels of traffic etc. Such data is hard to obtain from the existing services mostly due to legal and technical issues. And that is, if we talk about current state. Obtaining the data for a possible scenario is impossible from these services due to their nature. Having a way to create such transport models in an efficient manner would also help the company to address similar issues with other cities.

In Chapter 2, we focus on describing the theory behind modelling of travel demand, Chapter 3 focuses on the novel big data resources such as mobile networks or social media to figure out its potential for the activity based model. Having evaluated the possibilities, I list the ways to fill the gaps of big data usage in Chapter 4. In Chapter 5 the theory is applied to the Capital Region of the city of Helsinki, Finland to show how to combine the data sources together in order to get a viable model. To make the model even more useful, only open data and assumptions are used to create the model. In Chapter 6 the resulting model is discussed and the modelling process is evaluated. Furthermore, the possibilities for future development of the model are discussed and there is a reflection on the big data experience. In Chapter 7 the contributions of the thesis are briefly summarized.

2 Travel Demand Modelling Approaches

2.1 The Idea of Transport Modelling

The idea of modelling transport demand first arose in the 1950's as a means to aid building of the highway network in the US (Weiner, 1997). The model implemented a so-called "four step methodology", which consisted of trip generation, trip distribution, mode choice and trip assignment (to the final route). This was a major breakthrough back then, but it came with some major setbacks that are discussed later.

2.2 Categorization

The models could be categorized in multiple ways according to their scope and philosophy. One perspective is described below and illustrated in Figure 2.

2.3 Scales of Models

2.3.1 Microscopic

Microscopic models are trying to describe precisely the physical characteristics of individual moving objects (cars, bicycles, pedestrians) in the transport network. The interaction between the following vehicles are usually modelled using behavioral models, for example “car following model” in PTV Vissim. (PTV Group, 2014)

Moving objects are usually moving across links and nodes and for changing the lane a separate lane-changing model is used. Travel demand can be created with very high spatial and temporal precision (usually there are entry links or parking lots). These models can be denoted as “bottom-up”, since the result is built up from single actors interacting together.

2.3.2 Macroscopic

Macroscopic models model the transport network as a series of links and nodes. The whole intersection is usually represented as one node; the roads can be represented as two links, one for each direction.

One key difference from the microscopic model is that the travel demand generation is usually aggregated to transport analysis zones (TAZ), (McNally, 2000), even though some models might have finer spatial units such as facilities. These models belong to the “top-down” scheme since the results are derived from statistics without any interaction between actors (agents). Traffic participants are modelled as “streams” of vehicles rather than individually.

2.3.3 Mesoscopic

Mesoscopic models are somewhere between macroscopic and microscopic models. They preserve the links as in microscopic simulations (for example lanes on the intersection) however they will not usually recognize single vehicles. Another possibility is represented by models with single vehicles, but simplified intersections. In Matsim for example, the whole intersection is usually one node, neglecting all the intricacies happening inside it. Lane changing models could be non-existing.

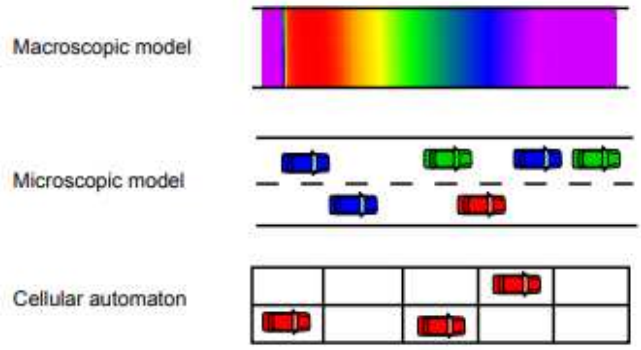


Figure 2 Illustration to different level of detail in modelling traffic flow from (Kesting, Treiber, & Helbing, 2008). Cellular automaton is an implementation of mesoscopic model.

2.4 Four Step Modelling

Four step modelling is a type of modelling that implements the macroscopic scale, transport analysis zones and a modelling process composed of four stages:

1. Trip Generation
2. Trip Distribution
3. Mode Choice
4. Trip Assignment

In Trip Generation the productions and attractions are generated for each zone. These should be proportionate to the amount of trips that start/end in the specified TAZ.

In Trip Distribution the Origin-Destination matrix is being created using the productions and attractions from Trip Generation. The amounts of trips in the matrix can be assigned by estimates from gravitational model, transportation survey or census. Gravitational model needs some resistance or travel impedance as an input, typically distance, travel time or travel costs.

In mode choice the trips need to have a travel mode assigned. In the simplest cases people would choose between car and public transportation, however more options can be added. As the choice set is always discrete the usual way to model the mode choice is use the Logit model.

The last phase, Trip Assignment, assigns a route to the trip. One of the ways to achieve that is looking for the user equilibrium, or assign the routes in such a way that no user gets an extra advantage by switching for an alternative route. This of course is a demanding process, thus a good technique is to assign trips by small groups until the full amount of trips is used. The catch about this method is that each individual sees the value of travel time differently in every occasion. (McNally, 2000)

2.5 The Critique of the Four Step Models

The critique for these models slightly overlaps with the Introduction part. Since the traffic flow was seen as key performance indicator in the past, the models were easily becoming a tool to expand the road network. The model would predict increased traffic flows and recommend building new capacity, thus indirectly increasing the options for motorists and inducing the demand even further. Later on, other factors started to be followed as well, such as accessibility, emissions, liveability or efficiency of the network (Litman, 2003). The models are not able to solve the transportation issues (te Brömmelstroet, 2017), however, they would make our decisions more informed. Without proper models, transport policies would be just pure gambling and traffic and transportation experts would have no numbers to justify their arguments (excluding surveys). The models should not dictate where new infrastructure should be built, as this is a political decision, but should predict in a somewhat reliable manner the outcome of each decision.

2.6 Activity Based Modelling

As an alternative to the previously mentioned four step model, the activity-based model has emerged. A fundamental premise of activity-based travel models is that travel demand derives from people's needs and desires to participate in activities. (Castiglione, Bradley, & Gliebe, 2015, p. 8) (Kitamura, 1988)

A strong critique was targeting one of the cores of the four step model, the transport analysis zones (TAZ). It was argued that such an aggregation means loss of key data (Mladenović MN, 2014) and thus a more detailed model is needed.

The second problem stems from the nature of trip modelling in the four step model as the trips are not connected to make a schedule of a virtual person, this is constraint that makes the activity based model more realistic (Castiglione, Bradley, & Gliebe, 2015, p. 1)

There are several frameworks for building activity based models such as in (Castiglione, Bradley, & Gliebe, 2015) or in Matsim (Horni, Nagel, & Axhausen, 2016). While the former still uses TAZ to simplify the modelling process, the latter relies on direct coordinates or facilities with coordinates and thus avoids the TAZ aggregation altogether. Therefore we will continue with Matsim to avoid the possible bias stemming from aggregation to TAZ.

2.7 Implementing Activity Based Modelling in Matsim

The activity-based model in Matsim needs these three specifications: the transport network, the virtual schedules of virtual agents in the model and a configuration file (Horni, Nagel, & Axhausen, 2016).

For Matsim the transport network can be obtained for free from openstreetmap.org (OSM), as the quality of the data is usually sufficient for such purposes, see (Barrington-Leigh & Millard-Ball, 2017). However, data should be at least visualized and filtered to check for possible mistakes. In theory, we can also obtain a network for public transportation, pedestrians, cyclists or other transport modes. To simulate the interaction of these modes, common links in the network should be used.

Traffic light information shall be possibly obtained from a separate source as OSM only gives an information about the location of traffic light, but not the implemented operation scheme. A workaround can be achieved by reducing the speed limits on the adjacent links accordingly, which might be difficult to calibrate.

A much trickier task is to obtain the virtual schedules / routines for people. There are more ways to achieve such a task, in principle they can be separated into the following categories:

- Survey the schedules as a whole (by travel diary survey for example) and expand them to the whole population
- Build the schedules using information about trips, people and facilities
- Build the schedules from population data, in a similar fashion as steps 1-3 from the four step model, then match the trips together

Any way one decides to go, one faces the dilemma of multidimensional task. In perfect model, the visit of the zoo is linked to the right person in the right time, in reality this presents a challenging task. If we count the dimensions for a precise problem, the problem equates to about 20 dimensions per person (personal data takes several dimensions, then each trip with its characteristics). Note that for simulation programs like Matsim, it is also necessary to provide information about the (initial) travel mode, further exacerbating the problem of getting the right data. Nevertheless, Matsim can also improve its people's routine if the right scoring for each trip is designed. That implies that Matsim enables you

to start just with passenger cars as travel mode and agents will change their habits gradually. Matsim could change/optimize all the activities, but to reach a meaningful model, at least compulsory activities should be assigned.

As in the case of the FSM, it is necessary to calibrate and validate the activity based model as well. For purposes of calibration, the traffic counts or mode split might be handy. For the validation the frameworks are being developed as well, one of the examples can be VALFRAM (Drchal, Čertický, & Jakob, 2015). They try to match the trips to O-D matrices and travel diaries on a statistical level. I am however using a more simplified framework comparing my results to hourly link volumes and public transport stop passenger daily volumes due to data availability.

2.8 Comparing the Resources Needed for Both Methodologies

One of the greatest advantages of the FSM is that it uses data that is mostly readily available or at least collected by the local administration. This means that only few extra surveys are usually needed, if any.

On the other hand, that data is not too useful for the activity-based model as it loses one of its greatest strengths, the level of detail. It is possible to build a model using these data, but the final output will probably not meet the requirement in terms of accuracy and precision. Check for an example of such a model built for Helsinki Region in (Väänänen, 2017) where despite using travel survey data the end result still misses the expected precision.

The strengths of the activity-based model are best met when an appropriate data source is used as a source of population's mobility. These datasets are on the other hand very hard to obtain and quite rarely open. In some cases, the local administrative authority or transportation organizer conducts a travel diary survey; in other cases, it is usually not available. The scale of the survey is critical for precision; a sufficient sample should target at least 4 % of the population (Nurul Habib, 2016).

Another idea comes from the mobile phone networks, since the providers have to record the position of the connected device within the network for billing purposes in a Call Data Records (CDR) format. These can serve theoretically as an essential source for Activity Based Models. However as described later, they are nearly impossible to obtain, at least in the raw state due to many factors mentioned in the chapter 3.

There have been also several other ideas to use so-called Big Data as a data source for the activity based models. They are discussed in the following chapter as well. However, none of them seems to reflect the reality as well as the mobile phone data or travel diary.

To conclude this part, gathering the data is one of the biggest obstacles for making accurate activity based models and one of the biggest reasons why four step models remain so popular despite their lack of detail.

Therefore, in the rest of the thesis we will focus on mapping the possible sources and combining them into fruitful mixes that may lead to building an accurate (or at least more accurate than at present) activity based model.

2.9 Data Requirements

Before we go through all the sources we need to know what exactly we are looking for and what issues may arise. For Matsim we need the network, the plans of the people and the public transport timetables. For the network, we would like to obtain the coordinates, capacity, allowed vehicles and allowed speed at least. For the public transport timetables, we need to get all the public transport schedules.

Then, for the plans, we need to get the population and its activities. In theory, if our plans contain close-to-perfect information, we do not need to bother with personal characteristics unless we desire to have them. Unfortunately, this is almost never the case and that means that we can break this dataset into two components - population and trips / tours. Just getting these separated presents a problem, as we need to find a way to glue them back together. In addition, getting trips just separately is not as useful since we need to organize them all into meaningful tours, usually with the same start and end point.

Population should include age, gender and might include nationality etc. where it is relevant. If people from a minority send their children into school where they are taught in the minority's language, it is a relevant information for further modelling. Age and gender are necessary in order to match the people with the right tours. If we for example know that only 5 % number of university students are over 30 years old, we want to get this number right as it increases the precision of the created model.

Tours are the second important component. Tour is defined as a series of trips originating and ending at the same place, usually home. If a tour is part of another bigger tour, it is called a subtour. Note that the difference from trip-chains is the cut-off such as 90 mins of activity performance (O'Fallon, 2003), thus a tour can be composed of multiple trip-chains. The reason why tours are in my opinion preferable to trips or trip-chains is the continuity. In FSM, it is theoretically possible that the person would travel from a suburb to the city to work, then go home and then go back to the city to pick up the children from school. However, if the time gap between the activities is very small, this is unlikely to happen and the person will perhaps try to chain the activities together. However, since tours are usually hard to obtain as they are in fact unique (especially the tours composed of more than six trips). I would even argue that long tours consisting of five or six trips could be used as a unique identifier of a person even with just the activity type and the district of the city. Thus getting the trips might still help if we have some hint how to construct sensible tours from them. For that, at least a decent temporal and spatial resolution is needed as well as a basic classification by the type of the activity.

3 Big Data

As Big Data recently became one of the important buzzwords (Waller, 2013), it started to lose its original precise meaning. While in computer science the term big data usually denotes something that is hard or impractical to process by traditional techniques (Ward, 2013), in the field of transport engineering it means just a large data set (Guido, Rogano, Vitale, Astarita, & Festa, 2017). Another comparison is that computer science definition usually uses the three V's, representing volume, velocity and variety (Ward, 2013). While the computer science definition of big data requires the product of these measures to be very high, in the transport modelling field it seems that just the velocity could be very low (once a year for example). The volume seems to be the key magnitude. The data like General Transit Feed Specification (GTFS), considered big data in (Guido, Rogano,

Vitale, Astarita, & Festa, 2017) would by far not fulfill the Intel's definition of big data mentioned in (Ward, 2013) that states the big data represents data flow over 300 TB per week. Unpacked GTFS data for Helsinki has just 500 MB and for the model, it is necessary to fetch it only once.

In my research, I am using the definition closer to the transport engineering field – simply data with large volume. Furthermore, in the Big Data chapter I only mention data that is passive and does not need an extra effort to be collected as it is usually a byproduct of some other service. Some big data is mentioned among traditional data such as census due to its static nature (collected once per 10 years) and use since the beginning of modelling.

I believe that the relation of Big Data and Activity Based Models can be quite beneficial as noted in (Anda, Erath, & and Fourie, 2017). In fact, provided that the mobile data would be openly available and in perfect quality, I could stop almost writing at this very row (though the data is only near-perfect). Nevertheless, maybe luckily for our privacy, it is not. Thus, one has to look for more imperfect data and analyze them. Can we learn travelling patterns from frequent tweets of a person? Can we learn the patterns from calls of the person? How about queries from the journey planner? I am trying to address these questions in the paragraphs below by going through all the possible big data sources.

Note that some big data is mentioned in the following chapter since they are not a new product (appeared before 2000) and since they are not obtained as a byproduct of some other service. Modern census might be something in between, since the data could nowadays be obtained from certain databases, however I believe in most cases this data is still enhanced by further surveys.

3.1 Possible Big Data Sources

Today's world offers many possible big data sources, I am going through several of them, the order roughly matches their importance.

3.1.1 Mobile Networks - Call Detail Records

Call Detail Records (CDR) are used by mobile network operators for billing purposes. They need to store the location each time a device makes a call, sends an SMS or uses mobile data (von Mörner, 2017). With the penetration rate of around 100 % worldwide (Iqbal, Choudhury, Wang, & González, 2014) we can also think that there are few people left out in this data which further reduces the bias.

The precision of the data is as good as how often the device is used and the precision of location determination depends on the density of base transceiver stations (BTS). In theory, it would possible to record the movement of the whole population throughout the day, calibrating only for those people who do not possess a mobile device. However, this precision obviously clashes with another society's value - privacy, therefore mobile network operators are required to store this data safely without exposing it to third parties, see the layers of privacy in (Morris, 2015).

However, researchers were able to get some mobile data for their purposes, and there are several ways to achieve that:

- The data is aggregated into OD matrices (case StreetLight Data Inc)
- The data is processed internally on the servers of the mobile network operator (case Smart Bay)

- Only a sample from the data is taken (case Barcelona)

Note: cases are described below.

Another case for sensitivity is if long-time data are given out or just the data for a single day. For the purpose of building a Matsim activity-based model of the city one typical day should be theoretically enough and perhaps even better than average travel patterns, since rare activity would tend to be lost.

3.1.2 Use Cases of CDR for Transportation Engineering

There are only few cases where mobile data was used to build an activity based model, as even the authors of the Matsim modelling environment concluded that the data is not available and used a synthetic data for their case (Zilske & Nagel, 2014). However, I managed to find a case where Matsim was used to build the model from mobile data in at least two cases.

The first case is project Smart Bay, located in California. The project obtained the data from the MNO AT&T and the data was processed internally on their servers (Pozdnukhov, 2016). The whole network has used 5 million agents, after being cleaned from useless data (for example where the records are very sparse). The authors calibrated and validated the data as well as proved that privacy is still protected. Privacy protection was not an easy task however. The model corresponds quite well to the reality and it proved to be even more accurate than the standard activity-based model following traditional modelling procedure as it was showing new development areas and the demand for travel of its inhabitants. This could not be achieved by the traditional process since the surveys would have to be conducted with much higher frequency, which is not viable financially. However, (Pozdnoukhov, 2015) has concluded that since activity-based models are already used and have very developed procedures in Metropolitan Transportation Commission in Bay Area in California, USA, it would be better to use this process as a complement to current modelling process. (Pozdnoukhov, 2015)

Another case where mobile data was used in combination with Matsim was the project Eunoia. Project Eunoia is funded by EU and aims to "*take advantage of smart city technologies and complex systems science to develop new models and tools empowering city governments and their citizens to design sustainable mobility policies." – Quote from Eunoia project website* - (Eunoia Project, 2012).

In the case of Barcelona, the project managed to obtain a mobile data sample of the size of 70 000 devices. This gave them a representative sample that is possible to scale up in quite an inexpensive way. The result was a good representation of the mobility of Barcelona's citizens. While writing this thesis the work was still in progress. (Picornell, Lenormand, Tugores, Dubernet, & Lucio, 2015)

There are also other use cases not directly related to activity-based models. One of them is Streetlight Data in the USA, where they gather data from a various set of mobile network operators and process the data on their servers. It is then possible to obtain data such as OD matrices or traffic flows, even from the past. Mining the data from the past is almost impossible in some cases, thus for some studies, these data might be invaluable. The official website lists several examples of usage of the data provided. (StreetLight Data, Inc., 2018) As the data is aggregated on secure servers, there are little issues with privacy, see description in (USA Patent No. US20150005007A1, 2015).

In Israel, the start-up called TrendIT used mobile data to offer some basis for marketing. Using the mobile data, it was possible to see how many people are located around your shop within a certain radius. This could be then used to optimize the performance of the business. However, the start-up failed to show profit and ended up being acquired by different company in 2017. The company's website - http://www.trendit.net/ - is not active anymore (by 2018) as well. See (Start-Up Nation Finder™, 2018) for details.

In Austria, the mobile network operator A1 wanted to provide mobile data for use in geomarketing (Der Standard, 2009) – in German. However, it stepped down from the idea due to privacy issues (Positium, LBS, 2014) and the website - **http://www.a1traffic.at** – is not available anymore (by 2018). However, it did continue using the data for research in transportation, exclusively with Technical University in Graz (Horn, Klampfl, Cik, & Reiter, 2014). Upon query, they were not willing to share the data for me to research.
The best case for mobile data usage seems to be Estonia, where the government is eager to get data in order to save resources from surveys. It is even planned that mobile data measurements will replace some traditional surveys (Positium, LBS, 2014). The data is provided through Positium Ltd company, which offers population data and OD matrices. The data was also used by many researchers, see again the (Positium, LBS, 2014, p. 43).

3.1.3 Social Networks

As unexpected as it might seem, social networks are another potential source of mobility data. They have a great advantage in that the data is usually open for developers, such as logs of single tweets on Twitter. However, it also comes with great hindrances, such as higher bias (one could argue that whole generations would be left out if their mobile phone penetration rates tend to be lower and much smaller sample of population. In some ways, it is similar to the CDR as we know the location only when the social network is being used and the post on the social network is geotagged. For Twitter, only about 16 % of tweets obtained were geotagged. This number was obtained experimentally by my own measurements of the tweet dump for Helsinki region; see *Appendix 1.* Excerpt of twitter datafor the example of the data, Figure 1 for tweets plotted on the map and *Appendix 2*. Code extracting tweets with coordinates and time, plotting them afterwards for the plotting script in Python. This data could give at least some insight into the life of the city, in some cases (evening activities) even better than what is stated in travel diaries (Chaniotakis, Antoniou, & Pereira, 2016).

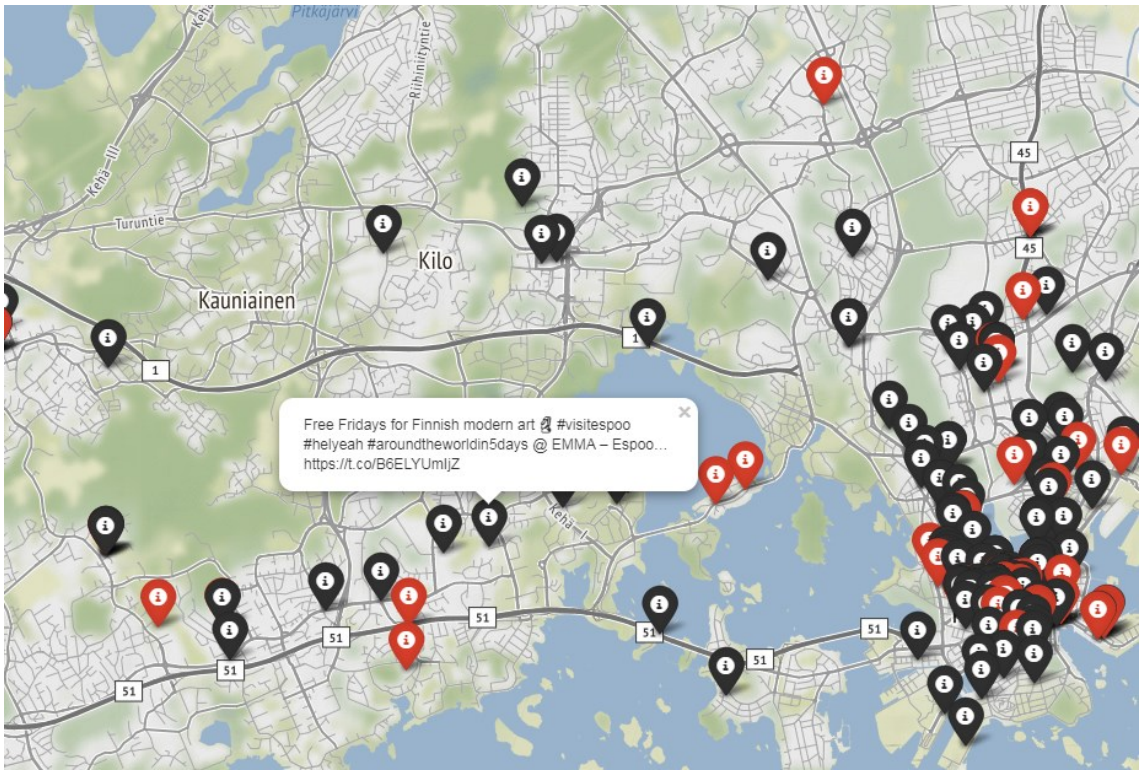


Figure 3 Geotagged tweets plotted on the map, colors represent different time period during the day

Apart from Twitter, there are other useful social networks as noted in (Chaniotakis, Antoniou, & Pereira, 2016, p. 68). The most useful seems to be Instagram, successfully used in (Di Minin, et al., 2016).

No cases of using social networks directly for building the activity-based model have been found in the studies. However, it is possible to find studies where this source has been in combination with others regarding accessibility and other related phenomena (Tenkanen, 2017).

3.1.4 Openstreetmaps

Openstreetmaps.org (OSM) is an open world map database created as an alternative to corporate-owned sites like Google Maps or Here maps. Not only it offers the map data but also the database model is open and very straightforward, composing of nodes, ways and relations. The network is almost complete, as tested for example in (Barrington-Leigh & Millard-Ball, 2017). The content is provided by users who voluntarily contribute to the database with their entries. As of now, automated filling of the database is rather discouraged as some users take editing the database as a hobby, see (OpenStreetMap Wiki, 2018).

The fact that the database is created voluntarily by users is also its great disadvantage. The consistency of the database is hampered by different perception of objects on the map. Where one user would see a one way, another one could see set of ways, while it might just represent a node for another one. This can be demonstrated in the various ways how public transport lines are recorded into the database. Differences also do occur in tagging. Where some users would tag the stop as platform, others would tag it as a bus stop.

Understanding these differences is a key to successful mining of important data from OSM. Thus, mapping all the ways in which data is recorded should be the priority before

the mining of the data starts. Mining the data for the activity-based model is then quite straightforward and there are several ways to do that. One is to download the extract in .pbf, .bzip or .xml format. Another way is to use software programmed in Java called Osmosis and extract directly the data that are needed.

From OSM, it is possible to extract a road network (an especially fast way is to use the Matsim plugin in JOSM, but only for small areas). If we also aim to model other transport modes, we could extract the correctly tagged ways and nodes through Osmosis or from raw extract.

Apart from the road network, it is possible to mine facilities and even locations of households, services and workplaces. However, OSM would not enable us to see the number of employees or inhabitants as these features are not mapped, despite some efforts like the population tag. However, it using some assumptions we could count the floor area of the building and thus get a good estimate for capacity. For services, it is even possible to get the opening times tag, though the penetration is by far not 100 %. This is very useful for setting time borders to activities in certain facilities.

What we could not obtain from OSM are the schedules of people or OD matrices, so it is necessary to combine OSM with other sources. However, there is a possibility to use published GPS tracks. These tracks seem to correlate with traffic on the road, the sample of these is extremely limited, but they might give a hint in some cases. An example can be seen in (Ježek, Jedlička, & Martolos, 2015).

3.1.5 Route planners

Route planners can serve as a potentially fruitful source of information. It could offer an insight to origin and destination of people's travel as well as the beginning and end times of activities. In addition, we do get access to the offered transport solution. The limitations are that the route planners could be unimodal and that traffic flows between origins and destinations need to be calibrated. One of the ways to achieve that could be to aggregate queries for travel into zones, calibrate the flows according to some other source (survey) and then disaggregate it again. This has been partially done in (Lappalainen, 2016).
Possible bias are that not everybody is using route planners as well as not every journey queried has been realized. I assume that people especially use the route planner for travelling to unfamiliar locations or for the journey, which involves multiple transfers.
For this thesis, I obtained data from Reittiopas (Helsinki Region journey planner) and I plan to investigate the usability of this data further in the thesis. However the data is not always available freely, for example I have not found a way to mine queries from Google Maps. Thus available route planners for data mining in queries would perhaps only consist of local transport providers and are not necessarily open.

3.1.6 Google Maps

Based on what information Google is able to provide to its users (Popular times for places, precise routing apps, traffic data etc.) we could assume that this data would be a perfect source for creating activity based models. Google gets its information from all the Android phones that give the permission to share the information. However, this data is the core of the advertising business of Google so it is not likely to expose this data to anyone else but the end users. Despite that, some information might be still mined using their API.

3.1.6.1 Places

API description: **https://developers.google.com/places/web-service/details**
Using the place ID as a gateway, Google provides address, coordinates, photos, price level, place and business' website among others.

3.1.6.2 Popular Times

Popular times are a special addition to places that is available to users, but not available through the API. The data is generated by users who agreed to share this information with the Google History Location service. (da Silva & Silva, 2018) As the name suggests, Popular Times reveal the frequency of usage at different times throughout the week. It has been proven by (Tafidis, 2018) that the information is reliable and can be used for travel demand modelling. Google however does not seem to expose this feature for developers as there are unfulfilled requests in Google issue tracker for more than three years. Still, searching on the internet still reveals some libraries to access these data, however with unclear licensing, so the usability of such library is minimal.

3.1.6.3 Route planner

Google Maps provide their own route planner that includes multiple modes, including car, public transport, walking, cycling or plane connections. It is worthwhile to note that Google stands behind the "GTFS revolution" that introduced a single standard for machine-readable timetables.

3.1.7 Mobile phone applications

The last option presented as big data source are mobile applications. It is possible to distribute these into two categories. One are mobility applications like TrafficSense developed at Aalto University (Rinne, 2018).This app records the journeys of the users with their permission and based on their journeys offers traffic information that is only related to the paths these users take. However, as the spread of this application started at Aalto just recently, it is not even possible to use that app as a representative sample for the Helsinki Metropolitan Area, not to think of other cities. Another type of these applications might be developed for the purpose of mobile travel diary survey; this is a great enhancement to the surveys relying on pen and paper as the participants might omit some of their destinations during the day.
The second category would be applications that are not primarily designed for mobility surveys but they do gather data about people's movement. The purpose to gather such data could be geomarketing or some other reason related to the location of the user. An example can be applications for taking geotagged photos or fitness applications like Strava. However it is again important to have the permission of the user to explore the data. For example Strava could be potentially used to map recreational usage of infrastructure that is usually missing (cycling volumes) – see (Jestico, Nelson, & Winters, 2016) and Figure 4 - and photo apps like Instagram could fill the blank spots in the leisure activities, see chapter Social Networks.

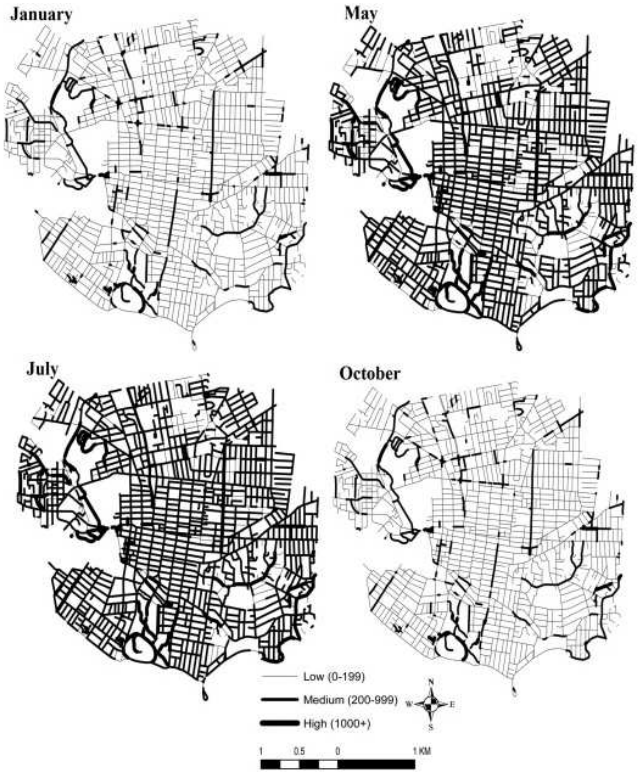


Figure 4 Cycling volumes in Victoria, Canada based on Strava application data (Jestico, Nelson, & Winters, 2016)

3.1.8 GTFS

GTFS, or General Transit Feed Specification is a great source for public transport time-tables. In theory there should be some relation between the offered capacity and traffic flows (or at least the maximum load section) thus it could help to validate the public transport traffic flows. It is also needed to model the public transport in the specified city. There are luckily ways to incorporate GTFS into Matsim, such as pt2matsim project initiated by (Poletti, 2017).

3.2 Comparison of the Existing Big Data Sources

In the following chapter, I would like to organize the possible sources for activity-based models and highlight their possibilities in a schematic way. In addition, I would like to propose some combination of sources based on the sources that are available. Please note that the scale is not based on any metrics and could be subjective in some cases such as processing difficulty. Furthermore, for most of the datasets those are just my estimates, as I did not get access to Mobile Apps and CDR datasets.

3.2.1 Evaluation Criteria

The criteria for evaluation were chosen according to the following table. The scale of rating the criteria is 1 to 5, 1 being the worst mark, 5 the best. In order to proceed as a suitable data source, openness needs to be five, as I will use the data commercially and other sources might bring licensing issues. High bias is tolerable if the dataset can be calibrated. The description of the criteria can be seen in Table 1.

Table 1 Evaluation criteria for big data sources

Criterion	Reason
Target	What data are we trying to mine
Openness	Reflects the accessibility of the data. Open data means you can download the content without any payments, sometimes registration might be required though. 5 means the data is completely open, 1 means it is very difficult to access the data. 3 means that the data might be accessible on special request.
Completeness	Reflects how big sample from the total number is available. 5 means (almost) complete population, 1 means only sparse records.
Bias	If the sample is focused mostly on a certain part of the population, it will suffer from bias. 5 means there is (almost) no bias meaning the sample is representative of the population, 1 means it is heavily biased.
Processing difficulty	Reflects how difficult it is (supposedly) to process the data. This scale is purely subjective according to the author. 5 means the processing is very easy and little programming or data mining skills are required.

3.2.2 Evaluated Datasets

I only evaluated those datasets that I managed to get access to or the datasets for which I believe I have enough information to evaluate (CDR, Mobile apps). The evaluation can be seen in Table 2.

Table 2 Big data sources evaluated according to the defined criteria, 1 being the worst, 5 being the best, see also Table 1.

Name	Target	Openness	Completeness	Bias	Processing difficulty
CDR	activities	1	5	5	3
CDR	transport mode	1	5	5	3
Social Networks Twitter	activities	5	2	2	3
OSM - GPS	transport mode	3	1	2	2
OSM - nodes	activities	5	4	1	4
OSM	network	5	5	5	4
Route planners' queries	transport mode	3	3	3	3
Route planners' queries	activities	3	3	3	2
Mobile Apps	activities	3	2	2	3
Mobile Apps	trajectories	3	3	4	3

3.3 Conclusion for Big Data Sources

For the network and timetables, big data is perfectly useful, however, for the plans of the people it proved less useful than expected, especially due to its (un-)openness. Working with the big data sources to get the plans of the population thus seems impractical, as I would need to compromise on the openness of my model, which I do not consider as beneficial. All the surveyed population data sources always suffer from a low score for at least one of the judged criteria. Thus, we need to have a look at the traditional data as well and combine them to get the quality of the plans that is desired. The idea is illustrated in Figure 5.

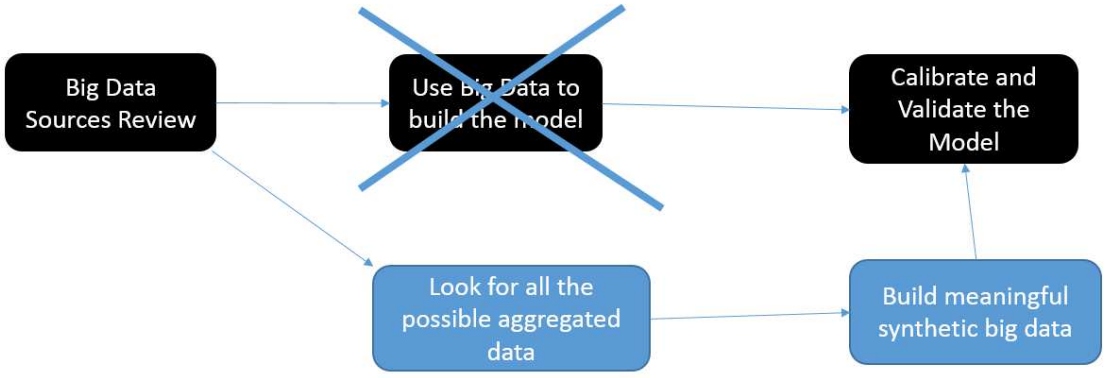


Figure 5 Changing the modelling process after the evaluation of big data sources

4 Traditional Data Sources

Traditional data sources shall include sources that were available before the year 2000 and they mostly rely on pen and paper or somewhat automated methods using people or cameras in the field to survey the population. A common denominator for this data is that it is laborious to obtain resulting in high costs, for household travel surveys the figure can go up to 350 $ per household in Australia (Stopher, 2007), other countries probably face similar costs. On the other hand, all of the big data sources mentioned above are obtained as a byproduct of some other service and thus with little additional costs.

4.1 Travel Diaries

The purpose of travel diaries is to record the travel behavior based on the daily routines of people. For the purposes of activity-based models, the survey should include type of activity, time, transport between activities and a detailed breakdown of the transportation connections used for the journeys.

If we could cover the whole population, travel diaries would be a great source to build the activity-based model, despite some imprecisions that occur when recording the data or that surveyed person would omit. However, there are two shortcomings of this survey; first of all, such approach would be extremely expensive, based on the previously mentioned figure for Australia, I would estimate about 200 € per household for Finland (converted from Australian dollar to Euro), resulting in 140 M€ cost for the full sample in Helsinki region. Second, not everybody would be willing to take part in such a survey. A reasonable compromise is to query only a sample of the people; however, that might already introduce sampling bias, especially for rare cases such as people commuting by taxi or an island having just 100 people being completely omitted.

Doubtable is the function of models built using these surveys as a ground truth for validation, as it was noticed that people fail to record all their trips as well as short stopovers, such as for shopping. (Jiang, Ferreira, & González, 2017, p. 9)

4.2 Transport Surveys

There are also various other transport surveys one can use to fill the activity based model, however they tend to be less suitable than the travel diaries due to their level of detail. One of these can be cordon survey (cordon representing an area boundary), where license plates are recorded to create origin destination (OD) matrix. These are however usually car-oriented so they would need to combined with a survey targeting other modes of transport as well.

Static traffic counts are another way to measure traffic. However, the greatest problem is that it measures traffic and not travel demand. Still, certain methods can be applied to estimate the OD matrices that could be used to build the model. (Van Zuylen, 1980)

Opposed to cordon survey, public transport survey can be used to complement these to get the holistic picture about origins and destinations of travels regardless of the transport mode.

However, in all these types of surveys the schedules of agents (virtual people in the simulation) need to be built using certain assumptions, which again introduces bias. Also, note that the schedules are the difference from the traditional four-step model, if the schedules are hard to obtain, it might better option to build the classical four-step model.

4.3 Population Census

Population census collects different data depending on the country. Usually only home and work activity locations are collected. In Finnish census, also the information about the recreational homes are collected. The problem might be period of the census, as it is usually done every 10 years (Statistics Finland, 2018). However, there are some countries like South Africa that conduct more holistic census that yields better data for Matsim. Still such data needed an extra enhancement through Multi-level iterative proportional fitting with an extra data source for more detailed information. (Joubert & Van Heerden, 2013)

4.4 General Statistics about Population

One of the powerful sources might also be the statistics, especially those that cover multiple dimension (for example location, age and work status). If only single dimensional statistics are available, then disaggregation methods can be used to recover the underlying data with certain level of precision. However, this data is often impossible to validate, so it might result in bigger error at the end of the modelling process.

An example of useful statistics are statistics about daily mobility of people, such as number of trips or modal split by travelled kilometers. However, one needs to be cautious when interpreting the data as it might ignore the whole groups of people (children under 7 years old or tourists) and might be outdated.

Another useful statistics is the time usage during the day, which is quite beneficial to get the activity durations and might as well include the time spent travelling. The activities might also be specified well into detail and include spatial information as in the case of Statfin.
The detailed data is however not open. (Statistics Finland, 2011)

In order to make sense of the population data one might also want to create households and families. Meaningful households might have positive effect on the precision of the simulation. Making synthetic households and families is a challenging process if the right data is not available, but certainly not impossible as shown below in the thesis.

Last but not least, the vehicle ownership and driving license ownership information might enhance the model quite significantly. In addition, statistics on ownership of the travel cards and passes are usually available.

There might be also further useful statistics for making even more precise model such as the data on disability, nationality or health might be handy.

The huge benefit of statistics is that they are usually open and the agencies publishing them need to make them for other purposes anyway, thus it does not come with additional

cost (apart from highly detailed statistics). The time spent searching for these might however end up being quite high, especially if there is also a language barrier. Different naming conventions of search phenomena also makes the research process harder.

4.5 GIS Data

GIS (Geographic Information Systems) data includes household locations, workplace locations, size of building, land use purposes etc.

They can be usually acquired from local administration and sometimes they could also offer alternative source for building the network from OSM. Some GIS data might cost money or is associated with specific software, which could end up in longer processing of the data.

4.6 Note on Elevation Data

Based on the examples in Matsim book, none of the scenarios mentioned using the elevation data in the simulations. However, this might have an effect on congestion and congestion propagation in the model, even on mesoscopic scale (Schönhof, 2007, p. 8).

Obtaining the elevation data is however the smallest problem, as it is possible to obtain it using Google Elevation API, Open Elevation API or from the national agencies (for example National Land Survey of Finland).

5 Test case – Helsinki, Finland

One thing is to talk about data sources and another is to test them by building the model itself. Since VTT had a demand for data that can be obtained from an activity based model of Helsinki if built, I decided this would present a suitable test case.

Helsinki is the capital of Finland located on its southern coast adjacent to the Baltic Sea. Speaking of Helsinki, it would not make sense to model the city isolated from other three municipalities that make up the Capital Region, see in Figure 6. The whole modelled area is thus composed of four municipalities totaling approximately 1,1 million inhabitants.

Figure 6 Municipalities of Capital Region (Pääkaupunkiseutu in Finnish) displayed in dark orange color by (RHYTK, 2018)

There are also some other factors speaking for Helsinki as a test case. It is a city of my master's studies (or Espoo, more precisely) and the city itself offers quite valuable data, although mostly aggregated into statistics.

For the final phases of modelling, I will use program called Matsim (Horni, Nagel, & Axhausen, 2016), written in Java. Since the program is written in Java, I will write my own extension for generating the inputs in Java as well.

5.1 Basic needs of the Capital region model

For my test case, it is desirable to set a couple of guidelines.
First, it is important to set the desired precision of the model. The model shall be based on real-world network featuring also links for public transport and bikes. As it is a mesoscopic model, pedestrian traffic will be simplified as teleportation (as opposite to routed modes) with the speed adjusted accordingly. Note that teleportation in the context of Matsim does not mean travelling with zero travel time, it means that the agent does not take part in the link queuing algorithms of the program.

The public transport will be loaded simply from GTFS, dead-heading will not be included. The loss of the precision might be relevant for some cases, but not for cases mentioned by VTT (passenger demand, traffic flow).

The biggest hurdle are the plans. The population component will have the size of the population of Helsinki Region (e. g. Espoo, Helsinki, Kauniainen, Vantaa). Each agent / person will have age, gender, role, family, household and will be located in a certain

building. The distribution of ages, gender and roles should match the one from statistics. In general, the composition of families should match with the available statistics as well as the household sizes. Household sizes should at least roughly fit to distribution of the sizes within the district. Population location should roughly match the population grid 250 x 250 m from 2017. Vehicles should be assigned to households or agents directly.

The model will have tours according to the distribution I am able to derive from the available data. Very long tours are initially excluded due to their complexity. Timeframe for trips should match the one from statistics. Popular times in facilities shall roughly match Popular Times of Facilities in Google Maps as they are considered as a reliable source (Tafidis, 2018).
The tours should be assigned to the people in a realistic way. For example, only a certain percentage of older students than 35 is acceptable.

Correct modal split shall be achieved by scoring (setting the correct utility parameters in Matsim) after a certain number of iterations (Matsim tries to achieve the user equilibrium by self-improving loop) within the Matsim framework and should match the modal split observed from statistics.

5.2 Loading the network

5.2.1 Basic Topology

The network in Matsim is composed of nodes and links, a common practice for transport models. Nodes usually represent a crossroad, although a crossroad can be composed of multiple “sub-crossroads”.

The nodes must have an ID and coordinates. These nodes are then connected by links. Those links have an ID, a node ID from which they start and finish (order is important). Normally, two links are requested for each part, one for each direction. Links also contain other important attributes such as number of lanes, free speed, capacity and quite surprisingly, length that can be virtual and does not need to reflect the distance between the coordinates. This measure is used to simplify the model and speed up the computation.

5.2.2 Format and Coordinate Systems

The only requirement for Matsim is that the file with nodes and links is in XML format and that the coordinates are in a Cartesian coordinate system, preferably with one meter being equal to one unit. This means that coordinates in WGS84 (World Geodetic System) need to be transformed into some Cartesian system such as EPSG:2393 for Helsinki. Due to the mathematical properties of transforming a spherical system into Cartesian, Matsim would need to be reprogrammed to simulate are larger than a country, since the distortions from the geographic projection would be too high. The Cartesian system is used due to its computation simplicity (Horni, Nagel, & Axhausen, 2016, p. 13).

5.2.3 Adjustments for Public Transportation and Cycling

To be able to figure out interactions, Matsim needs to have all traffic in one link. Links that run parallel in close distance or that cross each other do not have any interaction and are considered to be on a different level. Matsim allows specifying which links allow

public transport, bicycle or any other transport mode with programmed behavior by adding tag "modes" to each link. In theory, it is also possible to simulate routing of pedestrians, however due to the modelling complexity I will use the default of Matsim, pedestrians modelled by teleportation with time constraints.

5.2.4 Generating the network

The most common way is to generate the network from Openstreetmap as it already has quite good coverage (Barrington-Leigh & Millard-Ball, 2017). The easiest way is to download some already made excerpt for the area of study in the .osm format and then convert it into .xml format for Matsim using OsmNetworkReader class built into the Matsim core. Another way is to use the pt2matsim extension that helps to generate links public transport as well. The result can be seen in Figure 7.

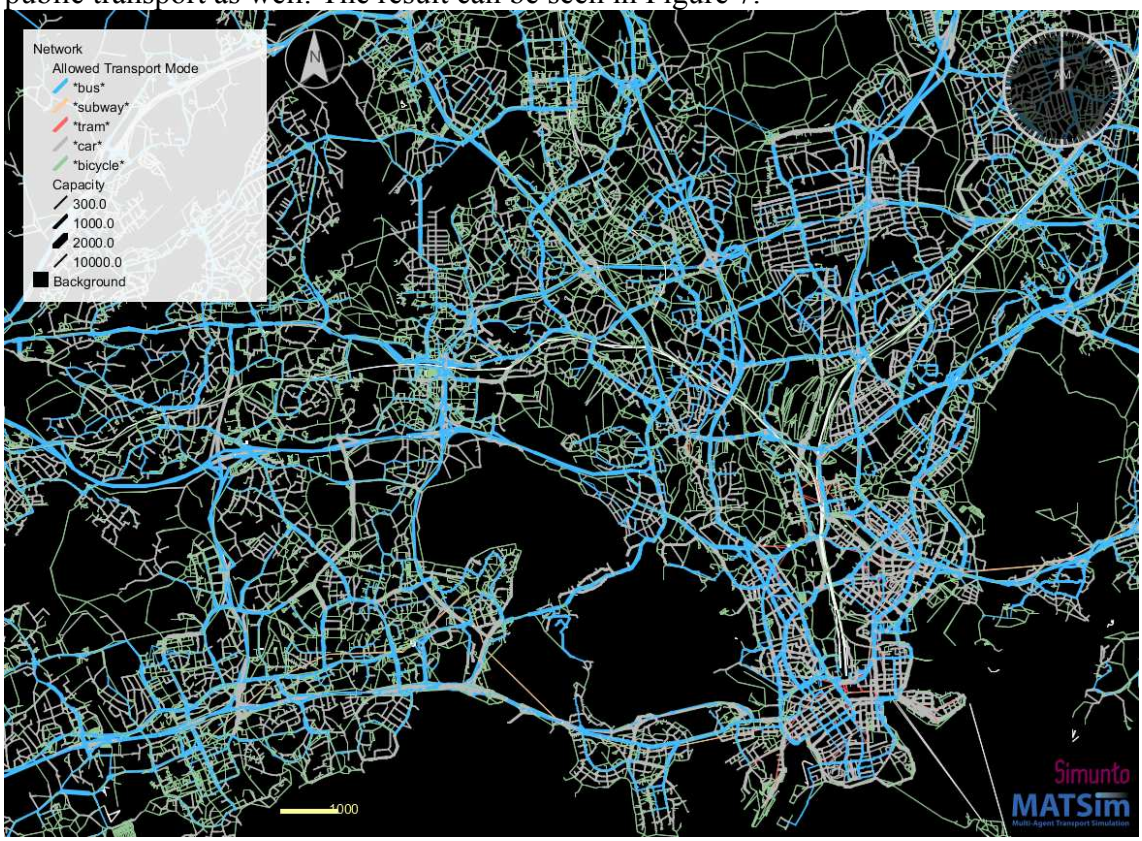


Figure 7 Snapshot of a network obtained from OSM, visualized in Via Simunto

5.3 Loading the Public Transportation

5.3.1 Public Transportation Implementation in Matsim

Public Transportation in Matsim composes of three parts - adjusted network, transit schedule and transit vehicles.

The adjusted network was mentioned in the previous chapter. Usually it is necessary to adjust the network for (trolley-)buses if the links are to be shared together. Metros are mostly running underground and will then have no interaction with overground traffic. Tricky cases are light rail or trams as they do interact with traffic. In OSM it is usually the cases that the interaction points are coded in as intersections, however it should be checked. The basic idea for trams should be that trams running separately have their own link while trams sharing the lane with cars should be sharing the link with cars or cars

should be allowed to take the tram link. Due to its time complexity, trams are loaded into my network, but attention to the interaction is omitted.

Next, one needs to have a transit schedule. This is provided as a separate file in .xml format divided into two parts. The first part describes all the stops of public transport. Each stop facility needs to have an ID, coordinates in coordinate system same as for the network and a reference for the link with which the stop is connected. Optionally one can specify if the stop has its own bay or its real name for a more understandable visualization.

The stops are connected by the lines. A line has the list of stops with assigned offset from the start for arrivals and departures. Then, links used by the route are specified as well as departures from the initial bus stop with the vehicle reference.

Finally, we specify the vehicles themselves. The identificators shall match the vehicle reference used in lines. This file is divided into two parts - vehicle type and list of vehicles. Vehicle type specifies the capacity, length and other parameters of the vehicle type. Note that one transport mode can have more vehicle types; however, transport mode is specified for the line, not for the type. Vehicles are just list of vehicles binding the vehicle reference and the vehicle type together.

5.3.2 Generating Public Transport Files

Generating such files manually for any city bigger than a tiny town would prove quite difficult. Luckily, there are sources and tools that help us achieve the same for bigger cities (semi-)automatically.

As a source one can use public transport schedules stored in the form of GTFS (General Transit Feed Specification). This format is used almost universally and includes almost all the information needed for Matsim. I used pt2matsim package (Poletti, 2017) that offers a way to convert the information about public transport as well as generate the network at the same time.

Another possible source is HAFAS format which I am not familiar with. OSM also offers resources for public transport; however, it has a major disadvantage in the form of missing timetables.

5.4 Generating the Population

In this case, I am using the term “generating” instead of “loading” for a purpose. The idea is that it is almost impossible to “load” a satisfactory population with its plans for a reason. The data needed can get extremely complex, depending on the degree of precision required.

5.4.1 What Do We Need to Create?

The idea can sound as simple as that - the people and their movements around the city. However, the devil lies in the details. How are we going to obtain the whole population? How do we achieve the same variety as we can see in the statistics? How do we connect the properties within the same persons? All these questions need to be handled in some way, otherwise creating a useful population is impossible.

And then the second part, how do we know that person A visits theatre B in the time C and stays for the duration D? We do not. And we should not. Nevertheless, we can give a realistic guess, such a guess that when we aggregate the data, it will make sense with the statistics.

5.4.2 Mobile Data as a Shortcut

In the beginning of my work, I planned to use mobile data to map the movements of people into my network, however as we have seen, it might come with bias, it has privacy issues and most importantly proved to be impossible to obtain for my thesis. I personally talked with the people in Finnish mobile operator Telia (special thanks to my supervisor, Claudio Roncoli, for connecting me with them) about the possibilities for my model. I realized from the talks, that what I initially planned (getting the mobile devices mapped on the network) was impossible due to privacy issues. I would not be able to touch the data I am using and the whole model would perhaps need to be on the servers of Telia and the data underlying the model could never leave Telia. Only aggregated data was possible such as OD matrix, perhaps enhanced by one-hour time resolution. This would have been very small improvement to the current state of the data that is available. Furthermore, it can come with some extra bias. One that comes to my mind is that if Telia is not physically present in some district (having a shop selling SIM cards) but other operators are, will it affect the data? Especially old people might still rely on this form.

Furthermore this story is not unique, even the creators of Matsim did not find any available mobile data for activity based model and they had to invent their own similar dataset to be even able to test them (Zilske & Nagel, 2014).
All this left me with a kind of frustration, as we could have a model built from a very granular data, but due to all these hardships mentioned, it will still be impossible for the next couple of years. In addition to that, with the new GDPR legislation introduced in the EU, it might be even harder to gather any data in the granular level and even statistics might become more difficult to collect.

5.4.3 Statistics

As a remedy, there is data without the burden of privacy and most of the time even freely available - statistics. The statistics themselves offer a generous amount of data about the people from the studied area. Furthermore, for Finland, many statistics are available at the level of districts, households or families. However, there is a major shortcoming - it is unconnected. If you were to generate the population with the basic indicators (age, gender, role in the society), you would have hard time finding a statistics connecting all these factors. By the role in the society, I mean child, student, unemployed, working or pensioner. Furthermore, we need to a place of residence to all the people with some desired precision (Based on the available data I set this to +- 250 m).

5.4.4 Synthetic Big Data

To bypass these issues I decided to make my own big data using several techniques. This big data shall fit the statistics and I do not claim that the data is connected in the right way. However, I tried to find the sources for the connections wherever possible, sometimes I needed to assume the connection. Such connection is sometimes tested by other statistics seeing the object from a different perspective.
Personally, I believe this should be a way to create the population for activity-based models as it has several advantages. First, most of the data is freely available (with citing the sources properly); second, you can start with a small amount of statistics and gradually

add more and more complexity until a satisfactory state is achieved. Third, since you have to guess the connection many times, you might discover previously unseen relations between your objects of study.

5.5 Implementing the Model

5.5.1 Generating the Network with Public Transport

I have started by setting up the environment for Matsim, which to me was not as trivial as expected. Having installed Eclipse I setup the project for pt2matsim, which is a special package to generate both the physical network and public transport network at the same time. This has some advantages as described in the chapter above. An idea of usage can be seen in Figure 8.

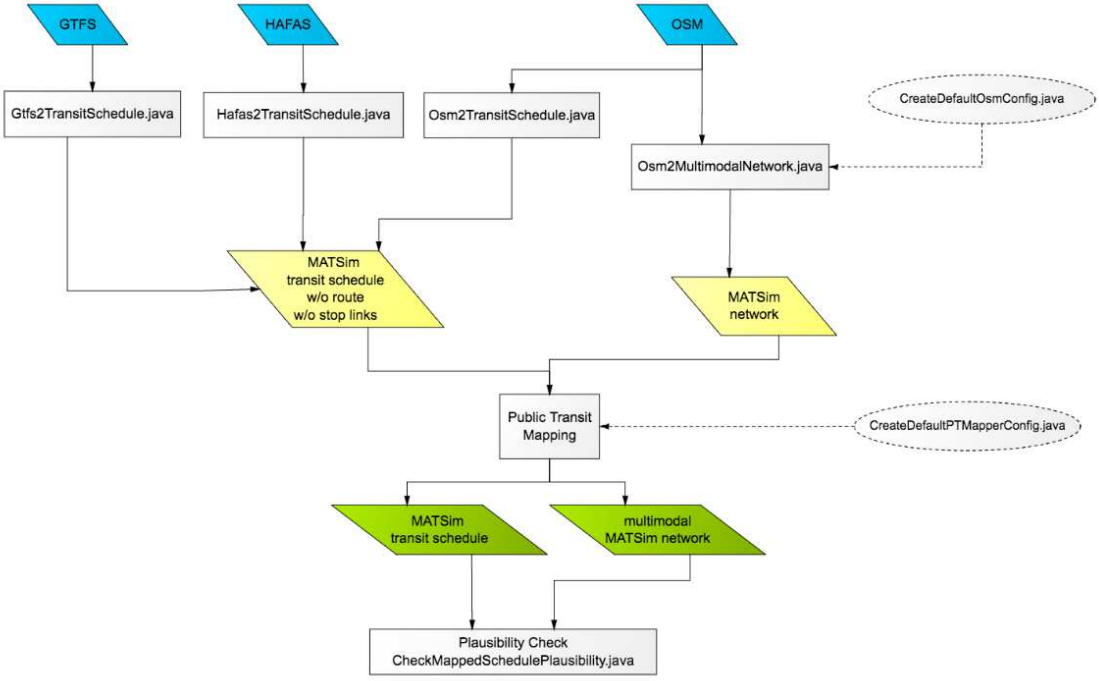


Figure 8 Flow chart for usage of pt2matsim package (Source: https://github.com/matsim-org/pt2matsim)

5.5.1.1 Links generated

First, I downloaded an extract of OSM from the bbbike.org. The raw osm format was chosen as it is required in the following steps. Using the Osm2MultimodalNetwork.java the file was converted into .xml file suitable for Matsim with the following extra: rail-based infrastructure was mapped as rail and bike paths were mapped as bike, see Table 3. Note that network modes might be more general than transport modes themselves.

Table 3 Transformation of OSM tags to allowed modes in Matsim network

OSM key and attribute	Mapped as
railway=subway	rail
railway=tram	rail
railway=rail	rail
highway (in general)	car
highway=cycleway	bike

5.5.1.2 Public transport schedule

The schedule was created from the GTFS. I used only one day of the schedule to make the process faster, and that is Wednesday 29/05/2018. This should be one of most typical days in the calendar as there are few holidays and university student are visiting the university, therefore suitable for the model. The following modes were mapped: bus, tram, subway, rail and ferry (to Suomenlinna). Once loaded the schedule has more than 8000 public transit stops.

5.5.1.3 Final network

The final network has around 150 MB and more than 100 000 links permitting cars. The bike network is more complex totaling over 300 000 links, thus appears to be more challenging for the router further in the thesis.

5.5.2 Generating the Population

In theory, one can skip generating the population and jump to trips provided we care only about trips/tours and we have enough information to do so. This, however, was not my case, so I tried to create a population with a desired level of detail.

The level of detail needed is always questionable. In my case, I was inspired by the report on the trips made by people in Helsinki region based on their age, role, location or gender (WSP Finland Oy, 2016). I thought it would be beneficial to have access to all these characteristics in my population. Furthermore, I found some studies related to households and families (HLJ, 2013), therefore, each person needs to belong to a family and a household. Note that there is a huge difference between the two as a household might contain multiple families. If we talk about families here, we talk about nuclear families as they are measured in (Statistics Finland, 2007).

5.5.2.1 Reconstructing the Population

As mentioned before, we need a population with a proper structure. If we go bottom-up, each person is a member of a family, each family is a part of a household, each household

resides in a building with an address and each building is part of a district, which composes a city and cities create the metropolitan area of Helsinki. For illustration, see Figure 8. Together we are speaking about generating more than one million people, even though just a sample might be used later.

From the previous paragraph, we can also tell what is not part of the model, which is social networks, family relations or work relations. I did not include them at this stage since I have limited period for my thesis, but I would see them as beneficial in the future. They might significantly affect the joint travelling, scheduling and other important aspects for the model.

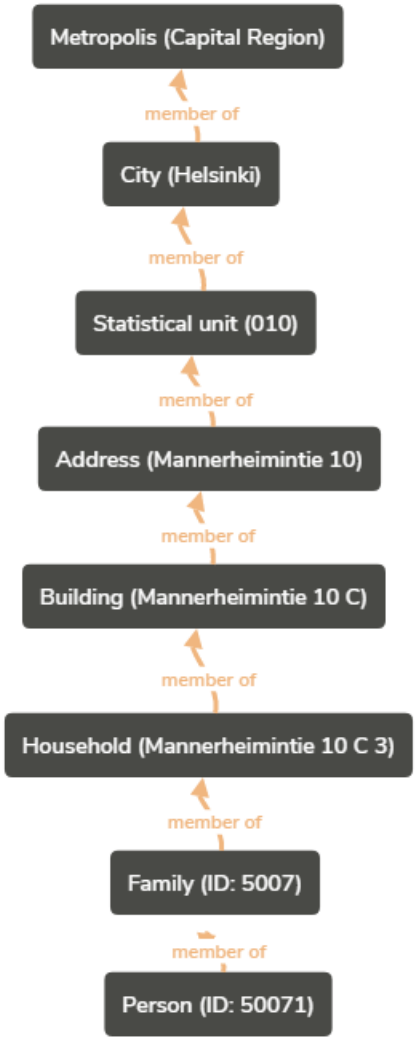


Figure 9 Scheme of idea of relations between objects

I had several ideas how to generate population. One that seemed promising was the genetic algorithm, where I encoded people as separate genes. The list of people did not change, just their position, therefore the children (genetic algorithm children) looked the very same when aggregated. The key was to find the right combination in terms of belonging to the right address, household and family. Despite the complexity of the scoring part, I managed to get the algorithm running. PMX algorithm was used to create new children (new combinations of genes), as this algorithm seemed to be the best for ordered lists (the whole set of genes was imagined to be an ordered list). Despite all that, the algorithm was less than successful, probably due to small variety in the offered solutions. The algorithm is not attached, as it has not been useful however it can be requested from the author.

Second and more fruitful idea was to fill the population top-down while keeping the desired characteristics. The algorithm works as follows:

1. Generate all the objects of population top-down until people are reached
2. Fix the possible errors by switching the position of people within family/household/district...

I discovered that it is beneficial to have the following levels of objects: city, district, building, household, family and person. Other combinations did not offer enough flexibility in terms of available statistics or were too difficult to create. In the end, I also added activities as a connection to the next part where I generate plans.

All the relations could be described as 1:n, so it is not possible for the object to be a member of multiple entities, thus a person can be a member of only a single family, thus family can reside in only one household, being part of one building that is located in one district. This idea was used since it significantly simplifies the generation of the population.

The generator is based on recursion therefore, each object triggers creation of its members inside its creation function. First, a city such as Vantaa is created as an object and then it is filled with its members. However, in some levels, it gets quite complex. For Location object for example, all the buildings are assigned with all their possible addresses in the initiation of the object and used afterwards.

5.5.2.2 Grid Cells and Statistical Units

Outside of the top-down hierarchy stands the population grid as well as the statistical units. Post-code zones (Statistics Finland, 2015) were also considered but ended up being redundant in the end and are only used later to correct the capacity for work in facilities. The reason why these two are not included directly in the hierarchy is that the availability of these might change from city to city and one of my aims was to make the program as universal and versatile as possible.
First, the grid is initialized for each city along with the buildings and addresses. The creation of buildings is described in the following part. Afterwards, the statistical units are initialized and serve later as a source to create households with the right size.

5.5.2.3 Buildings and Addresses

In order to locate the "home" activity of the simulated population the information about the buildings is necessary. Buildings in the model shall represent each building from the building register, however in this phase, they only represent buildings in which people can reside. As I was not able to find an open building register for Kauniainen online despite being able to do so for other cities in the capital region, I decided to use addresses to locate the buildings precisely. The idea goes as follows: All the addresses are loaded and assigned to the grid by the coordinates (both using the same coordinate reference system), then the buildings are created according the housing buildings number from the grid database (HSY, 2012). Finally, these buildings are randomly assigned to the addresses within the cell (permitting that one address is shared by two buildings, but quite unlikely, since there are usually more addresses than buildings).

5.5.2.4 Pre-loading the Age Distributions

As it turned out I needed to optimize the process of assigning the correct ages to the generated people and the best way to do that was to load the age distributions in the beginning, before starting the generation of households.

5.5.2.5 Household Generation

For households, the buildings are first assigned to the statistical units so that each building has one. Then, we are able to use these statistical units to generate the households as they come with the distributions of households by size. For that purpose, I used the dataset provided by all four municipalities that shows the household distribution by sizes, see Appendix 3. Households in Capital Region by Size and Statistical Unit. In this way, the households are generated while keeping their location with the precision to the statistical

unit, number and size distribution. This then indirectly means that the population numbers are fitting up to the level of the statistical unit. Note that there is one limitation as I am only generating households up to size 7 with regards of the difficulty of the next step. Also note that I had to adjust numbers for Espoo and Helsinki since they did not provide the distribution up to household size seven+, thus the numbers for the bigger households were extrapolated accordingly.

5.5.2.6 Creating Families

There are several critical steps bridging the gap between the households and their composition. As it turned out, the best way to cope with the unmanageable number of possible dimensions (age, role, gender etc. for up to seven people inside the household, and the people inside the household do have a relation between them, for example a household of 7 children would make little sense) was to use the nuclear families. Establishing the connection between the households, families and people we are able to reconstruct the population structure in a meaningful way, however not 100 % precisely.

First, we need to establish the link between household size and family composition. In the beginning I thought this relation is not measured at all, but Statistics Centre of Finland does offer them as a table called “P01D Asuntokunnat rakenteen ja henkilöluvun mukaan” (Households by composition and size) - (Statistics Finland, 2018). Since some municipalities included the table in their published materials, I was able to draw some inspiration on how the numbers should look like and designed my own table for the Capital region. I do use the same table for the whole region as I believe that the differences are minimal between municipalities. You can see the assigned composition likelihoods in the table below. Note that “a” represents a person living alone, “n” represents non-family person (a person that is not part of any other nuclear family) and numbers representing nuclear family as defined by Statfin with their respective size. Thus “32n” represents a household with 2 nuclear families of sizes three and two and a person without any nuclear family. That still does not exclude the possibility of them being all related between themselves. Also, if one person could belong to more than one nuclear family, the “younger” family is preferred in line with the classification by Statfin. The final assigning is seen in Table 4.

Table 4 Likelihood of household composition per household size

Household size	Family Composition	Cumulated likelihood of composition
1	a	1
2	2	0.923077
2	nn	1
3	3	0.974659
3	2n	0.998051
3	nnn	1
4	4	0.965583
4	3n	0.988528
4	22	0.992352
4	2nn	0.996176
4	nnnn	1
5	5	0.813953
5	4n	0.872093
5	32	0.968992
5	3nn	0.988372
5	2nnn	0.996124
5	nnnnn	1
6	6	0.679245
6	5n	0.754717
6	42	0.867925
6	4nn	0.90566
6	33	0.943396
6	32n	0.962264
6	3nnn	0.981132
6	222	1
6	22nn	1
6	2nnnn	1
6	nnnnnn	1
7	7	0.625
7	6n	0.666667
7	52	0.75
7	5nn	0.791667
7	43	0.875
7	42n	0.958333
7	4nnn	1
7	33n	1
7	322	1
7	32nn	1
7	3nnnn	1
7	222n	1
7	22nnn	1
7	2nnnnn	1
7	nnnnnnn	1

After that the matter gets even more complicated because now we need to design the whole family just based on their size. Luckily, such distributions already exist and are measured by Statfin. A lookup table with probabilities is used to generate the type of the family and then using the size and type the families are modelled starting from the mother. An age is picked for a mother (or childless woman) followed by age for her partner. The difference is modelled by normal distribution cut by some limitations (a very young woman can only make couple with a man of same age or older and vice versa). The age of woman (or a single father) is being picked with respect to the children they have according to the desired size of the family. This means that 20-year-old woman could not have 5 children in this scheme but can already have 2 or three. These limitations were obtained from the distribution of the age of mothers at birth, obtained from Statfin PX-Web Databases. A minimum age of mother is 15, maximum 55, while the likelihood of these is proportionately minimal.

Children are assigned to the families with children based on the age of mother (or single father). Father age is used as a limitation for the child's age, thus a couple consisting of 30-year old mother and 20-year old father can have 5-year old child at maximum.

The whole model does not take divorces into account for the sake of simplicity. However, I believe that the model is precise enough since it uses real age differences between mother and father and mother and child not forgetting the limitations of the difference between father and child.

After this step the people are generated, with their role in the family, age, gender, family number, household number, building number, address, grid cell, statistical unit and municipality. Some of these parameters can only be acquired through the membership inside another object. To simplify getting these properties a "virtual address" is being assigned to each person, making it easier to produce statistics on the created dataset.

5.5.2.7 Notes on the Algorithm

The algorithm relies heavily on the object-based structure which I was not too familiar with before I started writing this code. I am aware of the fact that more function should be made static as they load the same repetitively, but the aim of this thesis was to get a code creating the desired dataset and not creating the perfect code. For additional information it is possible to contact the author of the thesis.

5.5.3 Testing of the Generated Population

In the beginning of this chapter, we set a criterion for generating the population according to the available statistics.

5.5.3.1 Size of each city

First, we need to test that the size of the subpopulations roughly match with the size of each city within the Helsinki Region, results can be seen in Table 5.

Table 5 Data from generated dataset compared with the populations from statistics in (Statistics Finland, 2018)

Municipality	Population 2017 [Statfin]	Household dwellers 2017 [Statfin]	Population generated
Espoo	279 044	272 002	266 632
Helsinki	643 272	620 766	610 823
Kauniainen	9 624	9 361	8 805
Vantaa	223 027	218 293	214 035

Commentary or errors: The error is most likely generated between districts and households, since household sizes are slightly older (the data would fit between 2012 and 2014). However, the maximum error is around 10 % for Kauniainen comparing to the total population so I believe it is a fair approximation while keeping other characteristics.

5.5.3.2 Age Distribution

I believe that age distribution is best compared visually, thus I am showing the comparison in the following Figure 10.

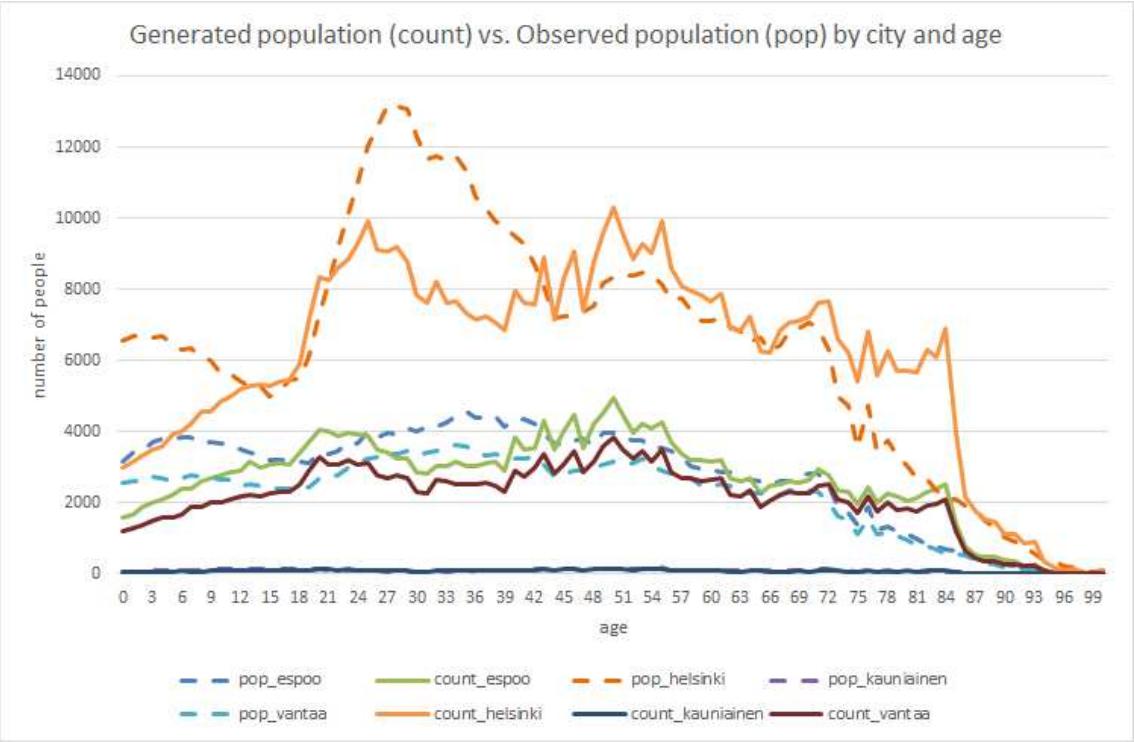


Figure 10 Generated population by age and city compared with the observed population from Statistics Finland

From the Figure 10 we can read there are some huge errors in three categories - children (age 0-12), young people (age 24-42) and old people (age 75-87). While the last category is overestimated, the first two are hugely underestimated. I tried to fix that problem by looking back into the algorithm and got the result visible from the following chart. It turned out that I did not need to handle childless families like a normal family, and mother age is newly drawn from the Statfin data - people by age, gender and role in the family.

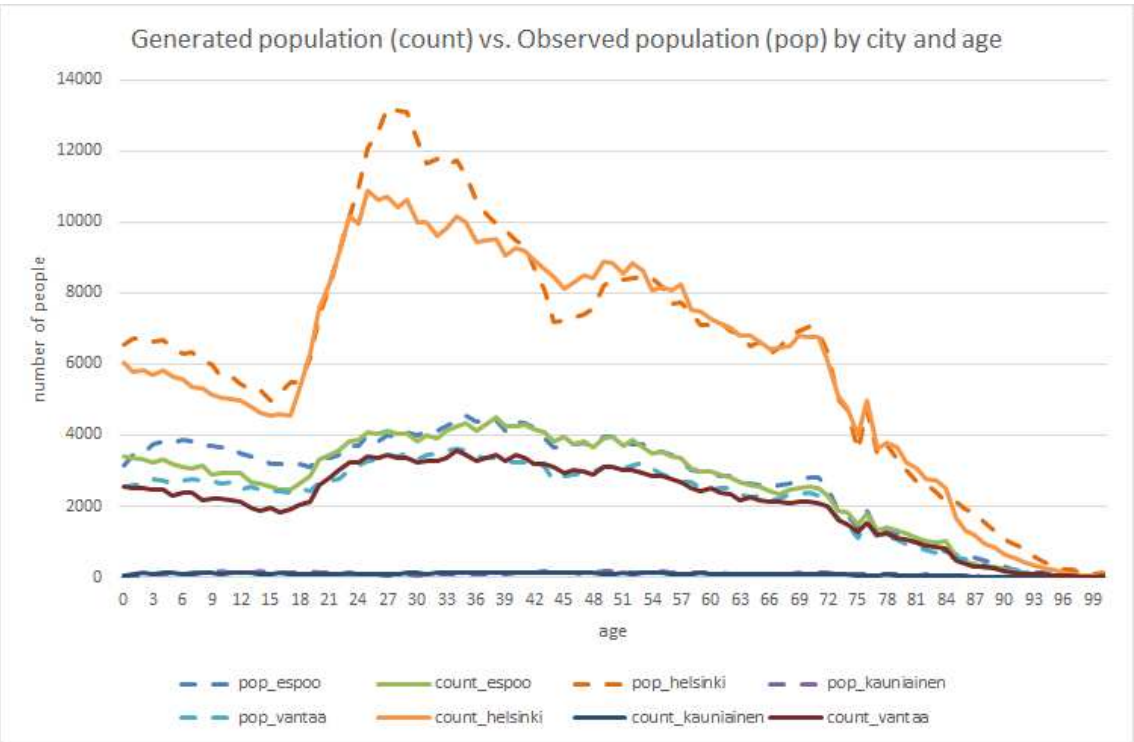


Figure 11 Generated population by age and city after correction

It seems that I managed to fix the gap for old people well and narrowed the gap for young people, yet it is not perfect. However, I believe the population is precise enough with differences mostly being within 20 % difference.

5.5.3.3 Gender Distribution

Gender Distribution means the ratio between men and women reflecting reality. Other possibilities are not considered since there is no available data on them from the statistics. The comparison with (Statistics Finland, 2018) can be seen in Table 6.

Table 6 Population by city and gender compared with statistics

Number	Men statistics	Men model	Women statistics	Women model
Espoo	138553	129694	140491	136939
Helsinki	305237	295686	338035	315058
Kauniainen	4654	4307	4970	4498
Vantaa	110613	103779	112414	110256

Commentary: From the table you can see that the differences are minimal apart from one case - women in Helsinki, where the difference is more than 20 thousand people. I do not have any good explanation for that difference, it could be perhaps caused by differences between the data used for each statistic. I did want to include gender since it has some effect on the number of trips and their categorization.

5.5.3.4 Family Status Comparison

In this case, the roles mean the family status within the nuclear family as defined by Statfin. There are 8 roles altogether, spouse (with/ without children), cohabiting (with/without children), single mother/father, child, living alone, living with non-family members and living in an institution. For the sake of simplicity, I aggregated the living alone, living with non-family members and living in an institution into one category. Due to the nature of this data, I believe it will make the most sense to use a table showing the difference from statistics, see Table **7**.

Table 7 Differences between the created dataset and statistics from (Statistics Finland, 2018)

Difference (statistics / model)	Alone/ Non-family	Spouse without children	Spouse with children	Cohabiting without children	Cohabiting with children	Single mother/father	Child
Espoo	24%	4%	-3%	5%	-6%	0%	0%
Helsinki	26%	1%	-7%	2%	-8%	-3%	-5%
Kauniainen	26%	15%	1%	20%	-2%	12%	3%
Vantaa	20%	5%	-4%	5%	-5%	0%	-2%

Commentary: Even though I managed to create some offset for the people not living with their families, I managed to be quite precise about the other categories. I might try to cover the mistake when sampling the population for traffic simulation.

5.5.3.5 Composition of Families

The composition of families reflects mostly the type and size of the generated families compared to the statistics. Please note again that we are talking about nuclear families as defined by Statfin.

From the table below (spans more than one page) you can see that the relative error is usually very low, and it only happens to increase when the number is generally low, therefore the relative error can be easily high. The family type is encoded as follows:

0 = single person, 1=married couple, no child, 2=married couple with at least one child, 3=cohabiting, no children, 4=cohabiting with children, 5=single mother with children, 6=single father with children. The precision is shown in Table 8.

Table 8 Differences between the generated families and their properties and statistics from (Statistics Finland, 2018)

City	Type	Size	Count	Statistics	Relative error
Espoo	0	1	51388	n/a	
Espoo	1	2	20711	21664	-4 %
Espoo	2	3	10492	9570	10 %
Espoo	2	4	12568	12271	2 %
Espoo	2	5	3756	3993	-6 %
Espoo	2	6	612	697	-12 %
Espoo	2	7	254	298	-15 %
Espoo	3	2	10139	10496	-3 %
Espoo	4	3	2944	2775	6 %
Espoo	4	4	2222	2177	2 %
Espoo	4	5	449	485	-7 %
Espoo	4	6	103	113	-9 %
Espoo	4	7	30	25	20 %
Espoo	5	2	4708	4984	-6 %
Espoo	5	3	3036	2758	10 %
Espoo	5	4	792	795	0 %
Espoo	5	5	204	208	-2 %
Espoo	5	6	53	52	2 %
Espoo	5	7	22	34	-35 %
Espoo	6	2	971	1052	-8 %
Espoo	6	3	433	370	17 %
Espoo	6	4	71	74	-4 %
Espoo	6	5	8	10	-20 %
Espoo	6	6	3	2	50 %
Espoo	6	7	1	1	0 %
Helsinki	0	1	175941	n/a	
Helsinki	1	2	46223	46759	-1 %
Helsinki	2	3	19439	17146	13 %
Helsinki	2	4	17928	17183	4 %
Helsinki	2	5	4981	4948	1 %
Helsinki	2	6	1017	1095	-7 %
Helsinki	2	7	554	562	-1 %
Helsinki	3	2	33163	33506	-1 %
Helsinki	4	3	7913	6805	16 %
Helsinki	4	4	4413	4108	7 %
Helsinki	4	5	752	798	-6 %
Helsinki	4	6	141	160	-12 %
Helsinki	4	7	73	59	24 %
Helsinki	5	2	13410	13652	-2 %
Helsinki	5	3	6688	5816	15 %
Helsinki	5	4	1650	1569	5 %
Helsinki	5	5	377	389	-3 %
Helsinki	5	6	136	140	-3 %
Helsinki	5	7	110	96	15 %
Helsinki	6	2	2514	2496	1 %
Helsinki	6	3	723	648	12 %
Helsinki	6	4	108	109	-1 %
Helsinki	6	5	14	17	-18 %
Helsinki	6	6	1	2	-50 %
Kauniainen	0	1	1418	n/a	
Kauniainen	1	2	816	919	-11 %
Kauniainen	2	3	323	331	-2 %

City	Type	Size	Count	Statistics	Relative error
Kauniainen	2	4	440	430	2 %
Kauniainen	2	5	186	226	-18 %
Kauniainen	2	6	33	33	0 %
Kauniainen	2	7	13	5	160 %
Kauniainen	3	2	201	222	-9 %
Kauniainen	4	3	64	58	10 %
Kauniainen	4	4	41	52	-21 %
Kauniainen	4	5	15	18	-17 %
Kauniainen	4	6	3	3	0 %
Kauniainen	4	7	3	1	200 %
Kauniainen	5	2	144	165	-13 %
Kauniainen	5	3	96	89	8 %
Kauniainen	5	4	43	38	13 %
Kauniainen	5	5	6	4	50 %
Kauniainen	5	6	1	2	-50 %
Kauniainen	6	2	28	37	-24 %
Kauniainen	6	3	21	17	24 %
Kauniainen	6	4	8	7	14 %
Vantaa	0	1	46438	n/a	
Vantaa	1	2	17324	18073	-4 %
Vantaa	2	3	7619	6982	9 %
Vantaa	2	4	8128	7870	3 %
Vantaa	2	5	2256	2416	-7 %
Vantaa	2	6	497	538	-8 %
Vantaa	2	7	224	247	-9 %
Vantaa	3	2	8827	9219	-4 %
Vantaa	4	3	3021	2772	9 %
Vantaa	4	4	2025	2008	1 %
Vantaa	4	5	421	449	-6 %
Vantaa	4	6	103	89	16 %
Vantaa	4	7	27	28	-4 %
Vantaa	5	2	4364	4574	-5 %
Vantaa	5	3	2680	2448	9 %
Vantaa	5	4	732	668	10 %
Vantaa	5	5	136	173	-21 %
Vantaa	5	6	50	51	-2 %
Vantaa	5	7	41	36	14 %
Vantaa	6	2	940	1020	-8 %
Vantaa	6	3	303	302	0 %
Vantaa	6	4	52	53	-2 %
Vantaa	6	5	11	10	10 %
Vantaa	6	6	1	1	0 %

Commentary: This table shows that the pattern of generated families according to their type, size and city is acceptable. Other characteristics such as age of the mother are not tested as I did not find any statistics to test them against.

5.5.3.6 Household Sizes by District

Household sizes must match to the district as the statistical district is the most detailed scale where I was able to obtain the sizes of the households.

Table 9 Comparison for household sizes between households and statistics, the match is 1:1 since the table was directly used to create the model

model	1	2	3	4	5	6	7	statistics	1	2	3	4	5	6	7
e111	1741	1160	317	248	77	16	7	e111	1741	1160	317	248	77	16	7
e112	1592	1165	409	297	97	20	9	e112	1592	1165	409	297	97	20	9
e113	1024	732	247	192	39	8	4	e113	1024	732	247	192	39	8	4
e114	289	354	214	207	93	20	9	e114	289	354	214	207	93	20	9
e115	34	117	69	90	43	9	4	e115	34	117	69	90	43	9	4
e116	223	253	118	163	41	9	4	e116	223	253	118	163	41	9	4
e117	254	304	143	188	80	17	7	e117	254	304	143	188	80	17	7

e118	793	616	202	176	68	14	6	e118	793	616	202	176	68	14	6
e131	1056	929	444	424	135	28	12	e131	1056	929	444	424	135	28	12

Commentary: From the Table 9 you can see that the results are exactly the same, that is because the numbers of households are directly based on these numbers. It also serves as a proof that they are not altered later in the algorithm by some erroneous code.

5.5.3.7 Population compared to the grid

The grid of 250x250 m squares obtained from Open Data by HSY (HSY, 2012) was the closest I could get to the number of people. Squares with zero people residing shall have no people in the final dataset. Some differences were expected as the household density is reflected only per district, not per cell. However, in general the data makes sense in terms of district densities.

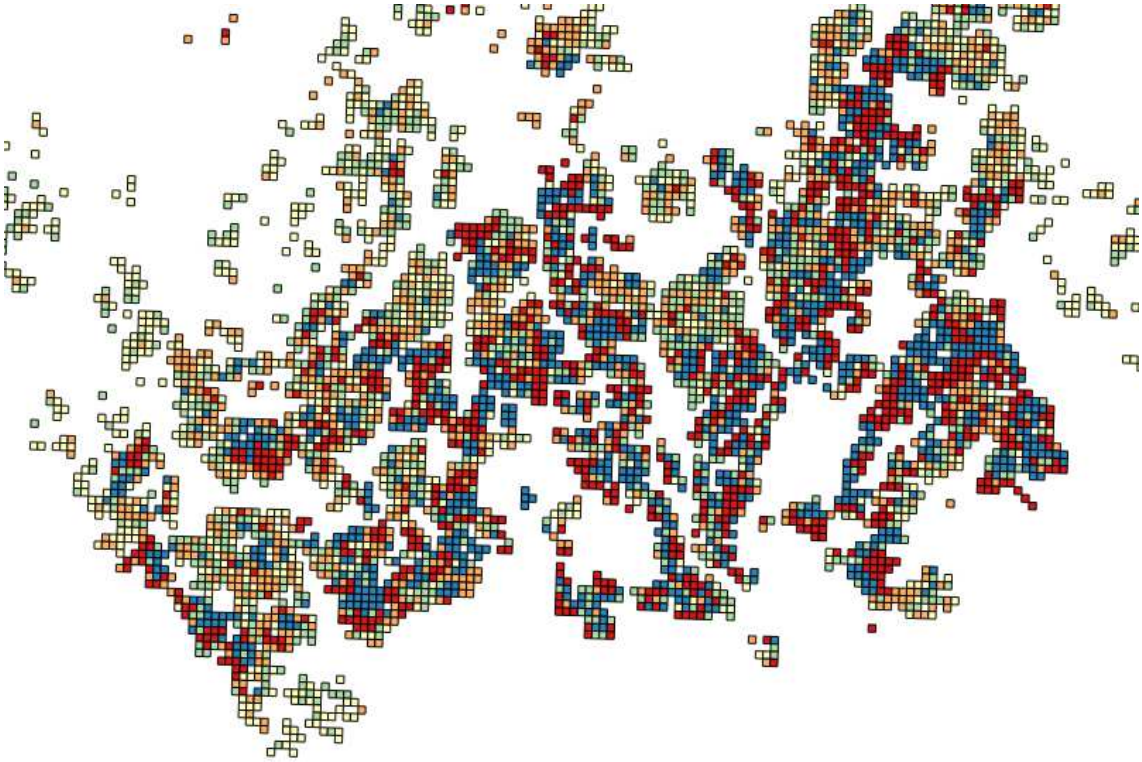


Figure 12 Population from the model compared to the population from the HSY grid. Population data from HSY Open data. Blue represents more people than in the population data, red symbolizes the opposite.

Commentary: From the Figure 12 we can see an expected picture. In terms of district, the population fits quite well, but in terms of small squares from HSY data they can be up to one kilometer away from where they should be. This would be possible to fix with a mutation algorithm that would gradually change the address of households within the district until it fits the picture. However, for this work one kilometer of error is acceptable.

5.5.3.8 Age Distribution on the Map

Even though the dataset is not designed to match the spatial distribution of each age group, I made a visual test of such fit. On the next figure, I present the grid showed on the heatmap in Figure 13 reflecting the location of people who are over 80 years old.

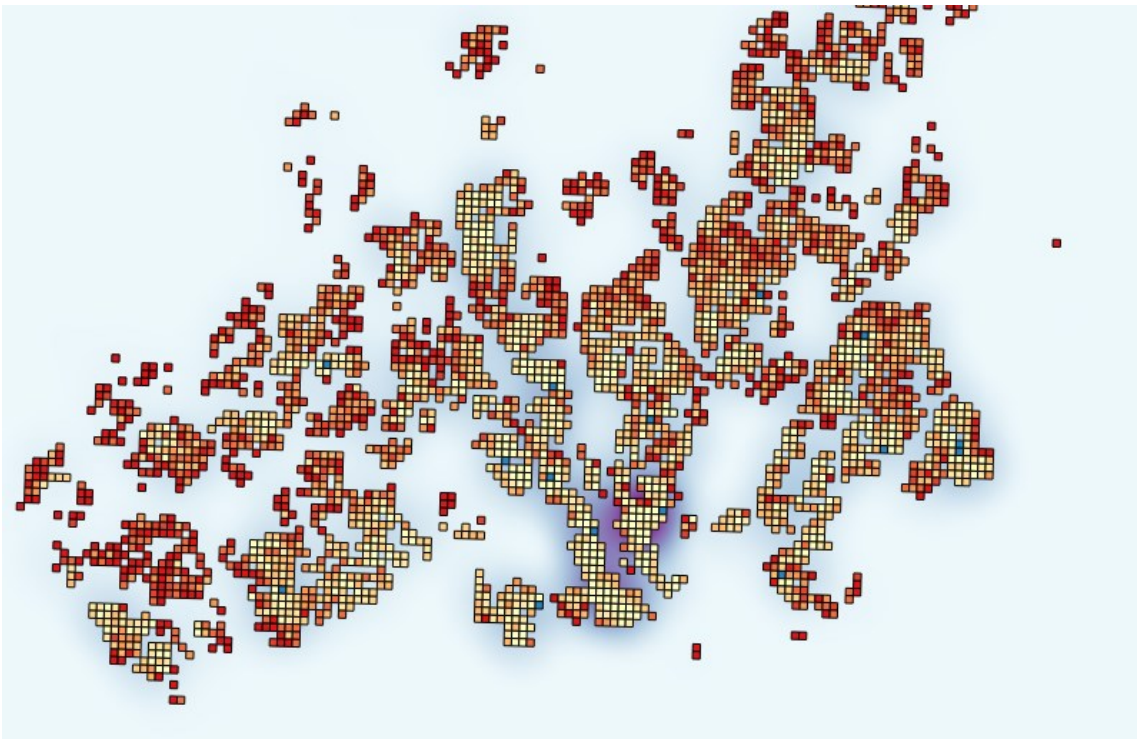


Figure 13 Grid showing people over 80 years old (white means higher density, red means low density) vs. heatmap of people over 80 years old from the generated population. Data from HSY or own dataset.

Commentary: Since the only driving factor to pick up ages is actually the household size (which actually matters more than I expected) we can see that the filtered population is located mostly where we would expect it. However, there is probably significant error in Kallio (approximately the area with violet color in Figure 13), where the generator gets confused since young people tend to have similar household sizes to older people.

5.5.4 General Commentary on Testing

With this population I tried to hunt many goals at the same time where I sometimes needed to figure out the relations just by myself. I managed to keep most of the desired characteristics within line however I am aware that it is not a 100% fit. But that was not even a goal, the goal was to have a population that would enable me to enter the next phase - trip planning and that was achieved.

5.5.5 Generating the Plans

Having generated sufficiently good population, we can move on to the next step, generating the plans for the people to move around. This will again require quite some creativity, but there are lots of measured data to give us a hint. The most useful data are the transport surveys, which measure the mobility of the people for the specified location with some degree of stratification, thus we will be able to use our generated population in all the possible layers to get as close to the real mobility as possible.

5.5.5.1 Requirements

As in the case of generating the population, it is beneficial to set up some criteria regarding the plan generation. First, we will need to enhance the people with some attributes like student status and work status. These will then restrict the behavior as we would

expect a link between this status and activities done. Membership in the family is already known to us so it might be used as a parameter as well.

At least to some extent, the characteristics of the people should be reflected in their mode choice. People who live in the remote suburbs should be more likely to choose car for their trips than people who live next to train station or people who live in the city center. Routes of the people are optimized so that people achieve user equilibrium - that is - a state where nobody gets an extra advantage by changing their route as in (Wardrop, 1952).

People should perform their activity in facilities. A facility can be quite abstract as a park or a statue within a park can act as a facility as well. Those facilities should have opening times if they are available and it would be optimal if they show similar load in time during the day as in Google Maps, as they tend to be quite reliable. (Tafidis, 2018)

Unfortunately, social networks and joint rides are out of scope for this thesis, as this phenomena seems to be rather complicated and would require even more preparation and literature review to implement.

5.5.5.2 Initial Data

Following quite deep review of the Finnish internet (in Finnish and English) I concluded that there is absolutely no raw data available as open data. This might be due to very strict privacy laws in Finland. To compare with, I was able to find raw open data for Torino, Italy within a weak with the help of Italian colleague. Even if I would perhaps get the data on request, I decided to use reports as my data source and use disaggregation techniques to reveal the data underneath. This has a great advantage that such a data should not be bound by any privacy requirements, yet it offers the level of detail needed. Of course, the precision suffers with each disaggregation technique, but after all it is a way of ensuring privacy for the disaggregated data. A data might incidentally describe real person's routine precisely, but the chance is very small.

Another advantage of the disaggregation technique is the scalability. While scaling up the data from surveys might come up with extra sampling bias (how do you scale the locations of activities?), disaggregated data works with the full sample from the very beginning, thus revealing the full possible variety within the disaggregating possibilities.

5.5.5.3 Tours in Capital region

Even before I created the dataset for population, I created a dataset of tours that are being made in the Helsinki Region (14 municipalities). Based on the available data I created roughly 1,8 million tours. (about 1,4 tours per person which seems to be a common average number in general).
The tours were quite simplified in this step so they would fit into the 21 most represented patterns as observed in the Microcensus 2000 made in Switzerland - see Figure 29. In this step I assumed partial transferability of these patterns to Helsinki region and calibrated them to fit with the Helsinki numbers. The generated tours were essential later to obtain hourly movements between the activity types.

Pattern	Percentage	trips per pattern
h-l-h	27.668%	2
h-w-h	26.342%	2
h-s-h	16.593%	2
h-e-h	12.147%	2
h-w-l-w-h	3.066%	4
h-l-l-h	2.438%	3
h-w-s-w-h	1.751%	4
h-s-l-h	1.581%	3
h-l-s-l-h	1.092%	4
h-w-w-h	1.760%	3
h-s-s-h	0.884%	3
h-l-s-h	0.803%	3
h-l-w-h	0.723%	3
h-w-l-h	0.992%	3
h-w-s-h	0.794%	3
h-e-l-h	0.406%	3
h-s-w-h	0.474%	3
h-e-e-h	0.214%	3
h-l-e-h	0.116%	3
h-w-e-h	0.087%	3
h-e-s-h	0.068%	3
total	100.000%	weighted avg = 2.23

Figure 14 Activity chain distributions as shown in (Balmer, 2007)

I did the calibration manually in MS Excel and the point was that the numbers must fit with the data from the HSL report (HLJ, 2013). The tours and absolute numbers can be seen in Table 10, the letters are explained in the previous Figure 14.

Table 10 Numbers of tour types as deciphered from (HLJ, 2013).

Tour	Number [thousands]
hlh	300
hwh	250
hsh	250
heh	170
hwlwh	60
hllh	144
hwswh	35
hslh	44
hslsh	30
hwwh	75
hssh	77
hlsh	7
hlwh	7
hwlh	25
hwsh	37
helh	30
hswh	5
heeh	2
hleh	17
hweh	9
hesh	15
hh	310
total	1899

5.5.5.4 Putting the Tours on the Timescale

Having obtained quite realistic distribution of tours I needed to distribute them in time. Again, I used the same HSL report as a hint. However, the report offered very few information on the chaining of trips in time and it even offered no way to tell the direction of the trip (from work or to work), thus I had to use certain assumptions (in the morning people go to work and vice versa in the evening), see the Figure 15 below.

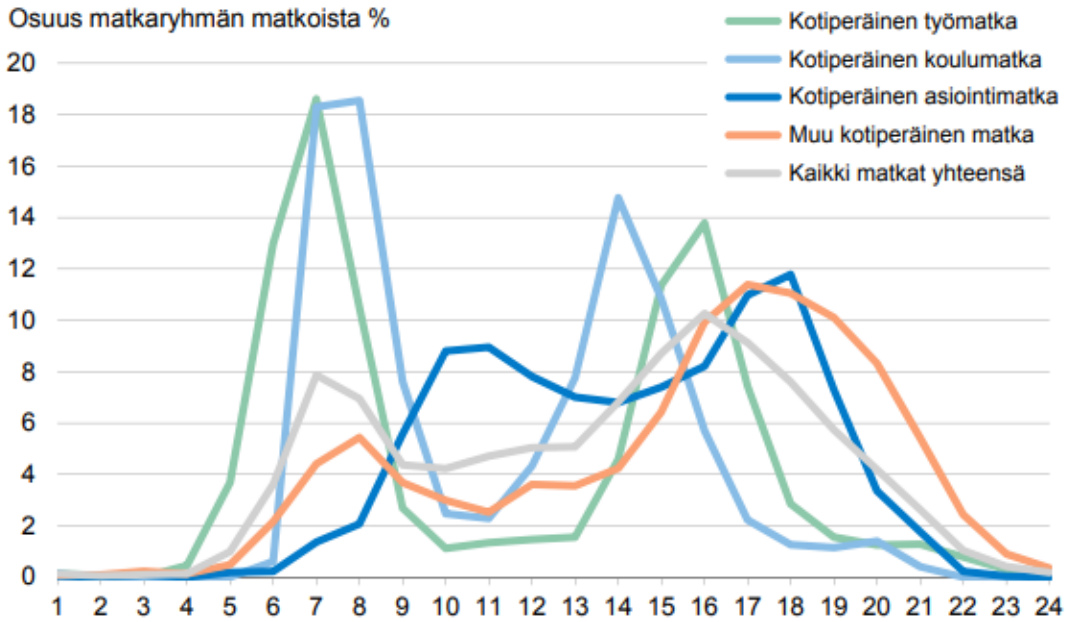


Figure 15 Workday trips by purpose and hour of departure. Green - homebased work-trip, light blue - homebased school-trip. Dark blue homebased errands-trip, orange other home-based trips. Grey all the other trips. (HLJ, 2013).

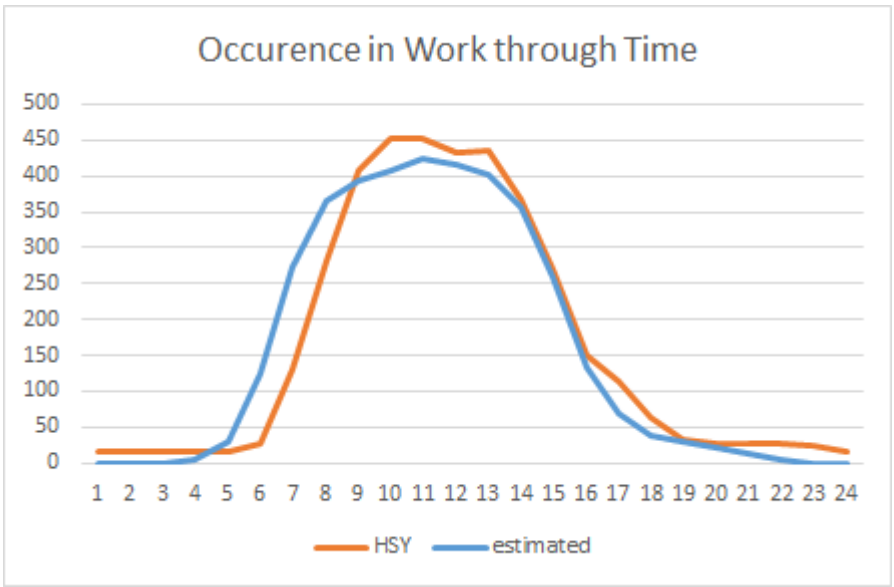


Figure 16 Occurence in work as compared with the HSY emission model

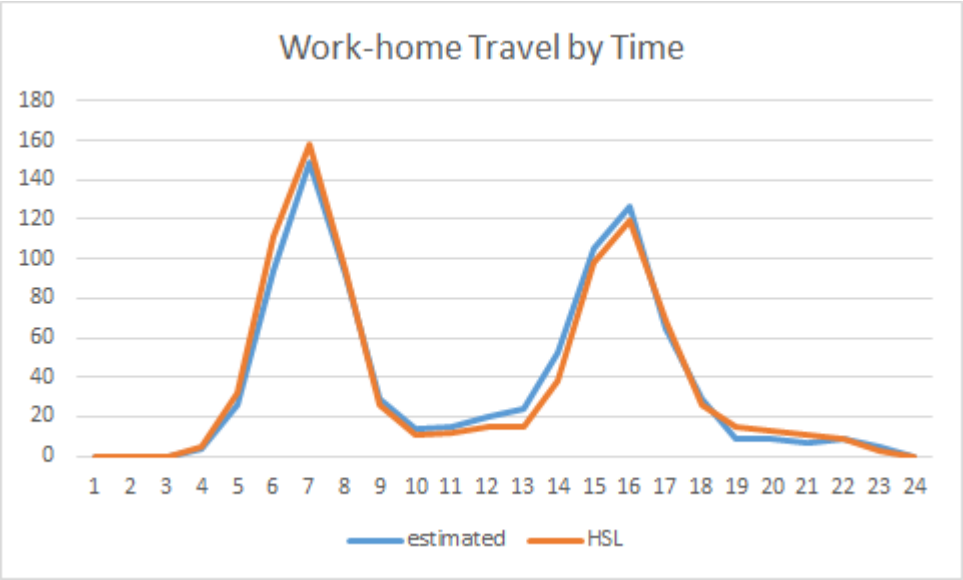


Figure 17 Travelling between home and work compared with the HSL study

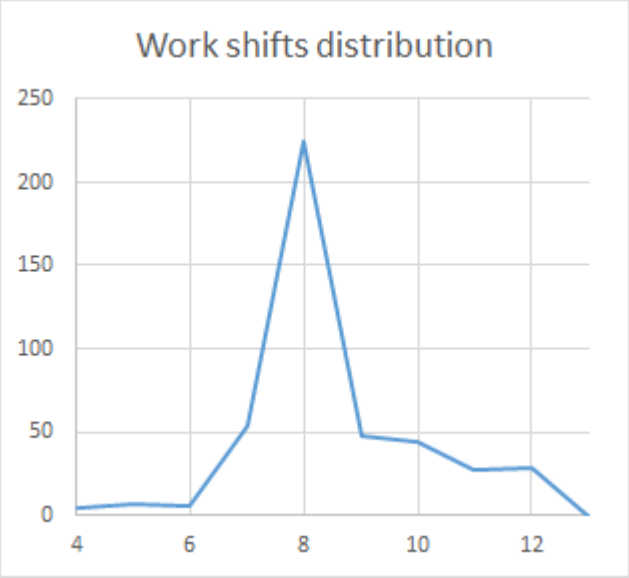


Figure 18 Work activity durations as estimated for hwh tours. I did not find any data to compare with.

The other simple tours were derived in a similar manner. The bigger challenge were the combined tours (for example hwlh) as there are too many factors to estimate (24^3 cells) so I used the timing from the simple tours and for the non-home based activities I used my own estimates. This enabled me to set the flows between activities throughout the time, an example can be seen in Figure 19 below. Also, I compared the estimates with HSY model of occurrence of people in time (excluding people in traffic) as seen in Figure 20 compared HSY Figure 21. The differences are caused by different classification of state of the population, for example all traffic activity is forced to the next activity and "muu" is split more in detail.

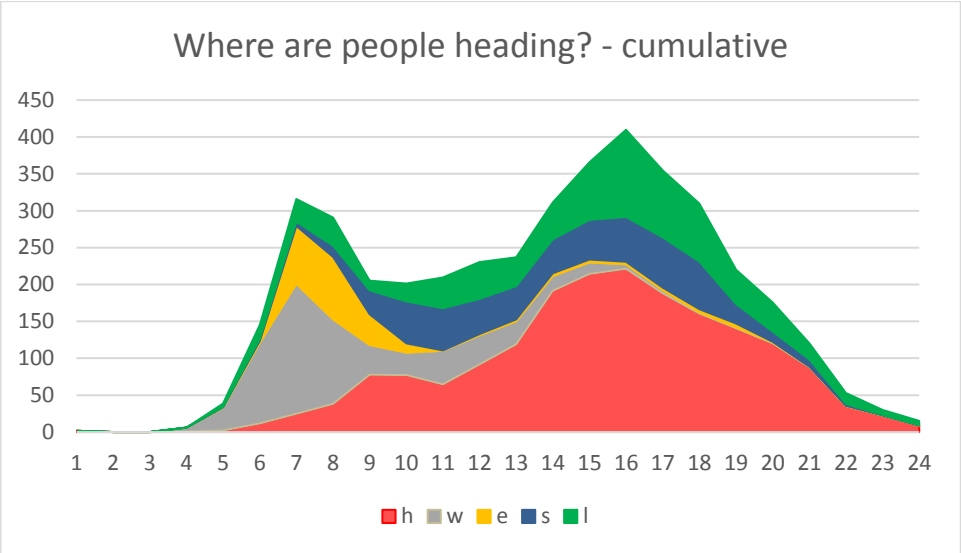


Figure 19 Trips in time by their destination activity, data deciphered from (HLJ, 2013)

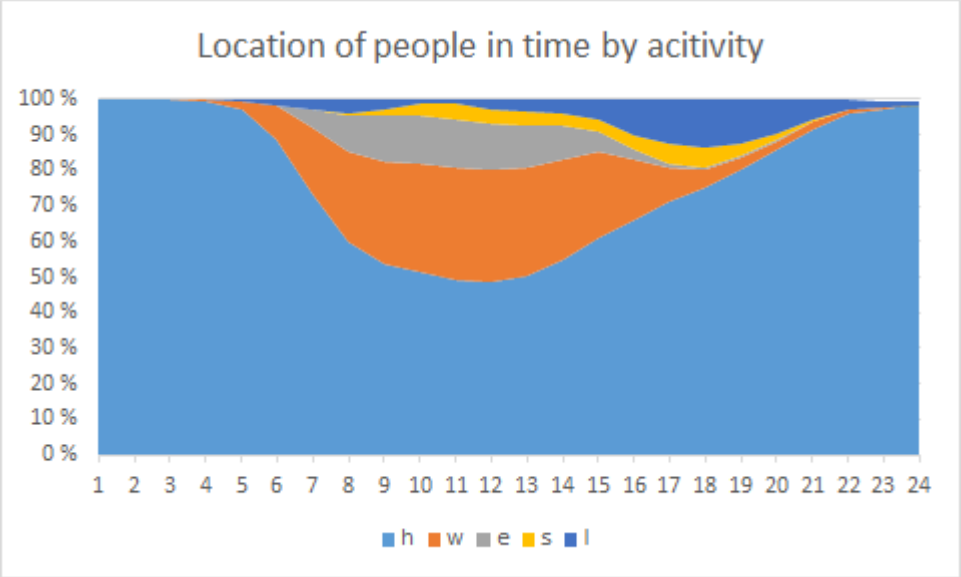


Figure 20 People from HSL region in time by activity they are performing (transport time aggregated to the following activity)

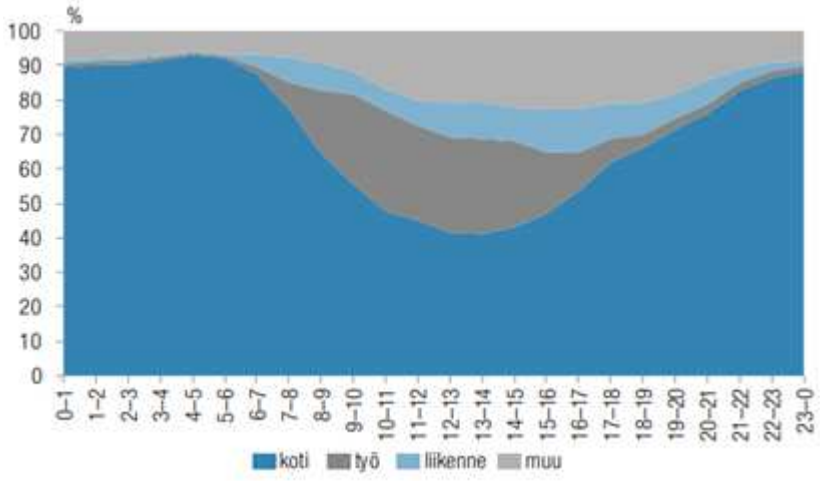


Figure 21 Figure of occurence of people in time by activity from the HSY study. Koti = home, työ = work, liikenne = in traffic, muu = else (errands and free-time activities in this case) – From (Kousa, et al., 2015, p. 170)

I tested the generated tours with the activity locations estimated from the capital region building register dataset (HSY, 2012). In this case I assumed equal space distribution by activity (for example about 30 m^2 per student) and distributed the activities randomly throughout the capital region. The results looked promising in terms of the pulses expected throughout the time as seen the Figure 22 below.

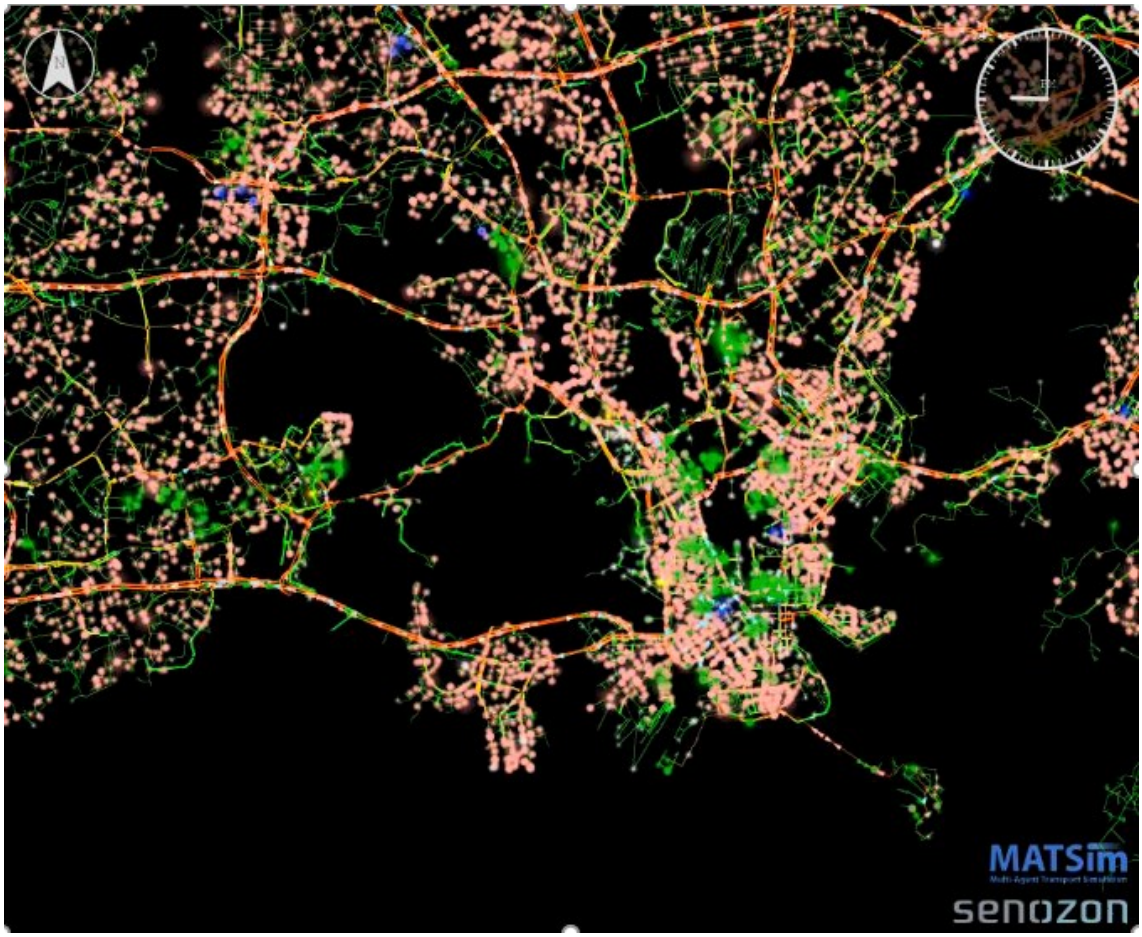


Figure 22 Helsinki in the evening - 9 pm. Green = leisure activity, Red = home activity, blue = shopping activity. Activity that just started to be performed is highlighted (more visible). Source: own presentation

Even though the testing showed that the distribution of activities in time seems to be (at least visually) realistic, it still did not generate realistic routines for the agents as in this case, every tour was assigned to an extra agent resulting in about 1,4 times more agents than in reality. There are also other errors associated with this type of simplification, as for example in reality the two activities that would be performed sequentially by one agent in reality can end up being performed in parallel by multiple agents. Also, the data for routines has value of its own, as we could then compare the resulting model with available statistics (for example daily travelled distance per person).

Thus, I decided to go one step further and generate full routines for all agents.

5.5.5.5 Getting the Mobility Patterns

Joining tours together in a meaningful way turned out to be an impossible task for me nevertheless I figured out a different approach. Since now we have all the trips between all the activity types distributed throughout the day by hour, we could assume that we observe a Markov process that is bound by transition probability matrices. This matrix

changes over the time of the day, otherwise it would be impossible to state when people start leaving their homes. A good illustration of the idea can be seen in (Chiba, Hino, Akaho, & Murata, 2017).

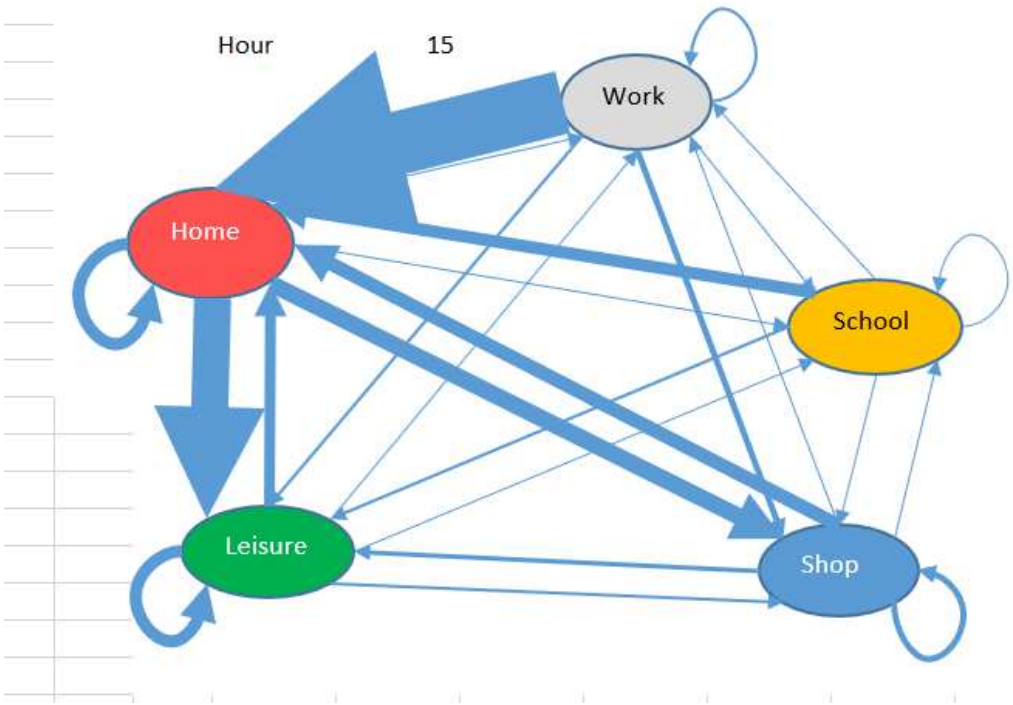


Figure 23 The idea for the transition matrix representing flows of the people between activities

The motivation for the idea is the following. If I tried to join the tours together I would perhaps get stuck by the infinite possibilities it creates. Furthermore, strange cases such as "study tour" and "work tour" together would be hard to eliminate (we only want to keep very small portion of these) and quite difficult evaluation criteria would need to be set up.

The time step model, on the other hand, makes all the variety possible while sticking with the original numbers for trips between activities and keeping the occurrence of people in activities in line the numbers by HSY. These numbers set the restrictions for the people to choose their routine. For example, once the agent chooses to go to work at 8:00, it will be very likely to stay there until 16:00, but other "paths" are also possible as long as there is "space".

The whole idea can be translated into graph composed of nodes and links. There is a node for every activity and every hour in the day, thus, since we use 5 activities, we get 120 nodes. Every node has 6 ingoing links and 6 outgoing links. 1 of the 6 links represents the agent carrying on the activity and 1 represents "restarting" the activity, e.g. home-to-home trip or work-to-work. Home-to-home is a special case since the agent is making a little subtour leaving and arriving at the same place. For other activities, it might happen to be the same, but more likely it is not. This might cause some issues, where agents might swap workplaces during the day. However, since there are also business trips that are not explicitly modelled, it is completely wrong that some agents visit more workplaces during the day.
Every link and every node have assigned a capacity. As in Kirchhoff's law, the input and output of the node must be equal. The capacity of the node represents the occurrence from the HSY study while the capacity of the link represents the number of the trips as derived from HSL report. However, the number of people staying in the link must be computed,

which is hard, since we can well estimate the number of people in the activities in the night. Then since we know the inflow for the night nodes and the outflow to activities (trips from HSL) we can easily get the number of people staying at the activity.
The whole idea is shown in the following Figure 24.

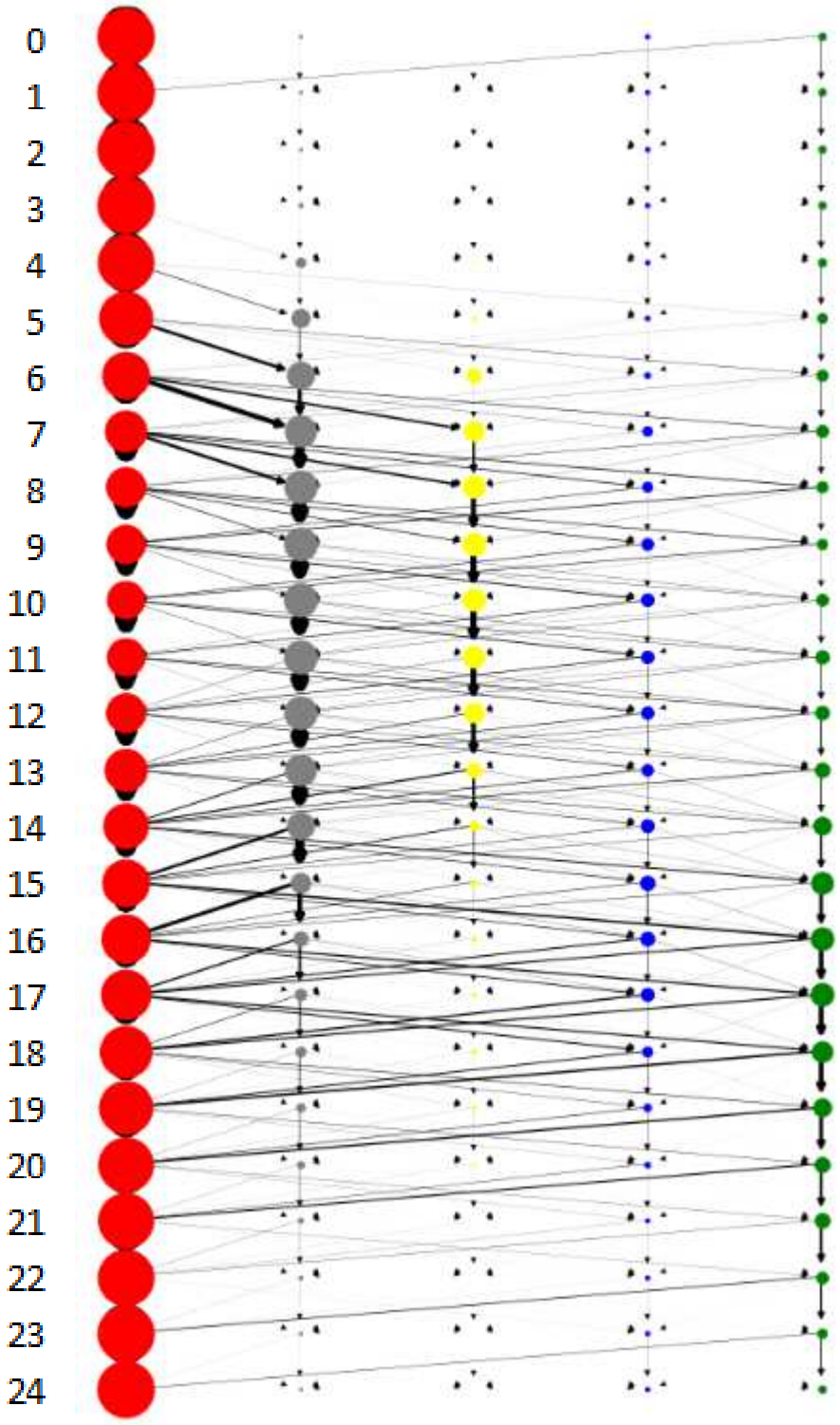


Figure 24 Time step model, vertical axis represents the time of the day, horizontal activity, parallel edges are merged into one. Colors: red=home, grey = work, yellow = education, green = leisure, blue = shopping/errands

Once we established the graph we can run a simple script where the randomly picked agents choose their next node in random order respecting the capacity of the links. In this way we assure keeping in line with the original numbers, yet we get all the diversity restricted in a reasonable way. So, it is possible for an agent to work and study at the same day, but it is less likely, and most importantly we don't have an overlap of activities in terms of time.

Also, it is easily possible to scale the number of generated patterns. I generated 1400 patterns, but one can also get thousand times more if the capacities are scaled accordingly.

This way still may some shortcomings though, as there is no way to ensure that a person will not be working the whole day, it is just quite unlikely (but it does occur). A remedy could be that agents would have a preference which link they want to choose next, however it might also worsen the "last agent problem" as the last agent would need to pick up the last remaining capacity. Another approach might be to run a mutation algorithm to swap subpaths where both agents have the same starting and ending node. This might lead to more reasonable total durations of activities for the whole, however, this approach has not been tested.

The obtained results proved to be in line with my expectations as you might see from the snippet below in Figure 25. The travel patterns shown in the figure seem to reflect the possible scenarios.

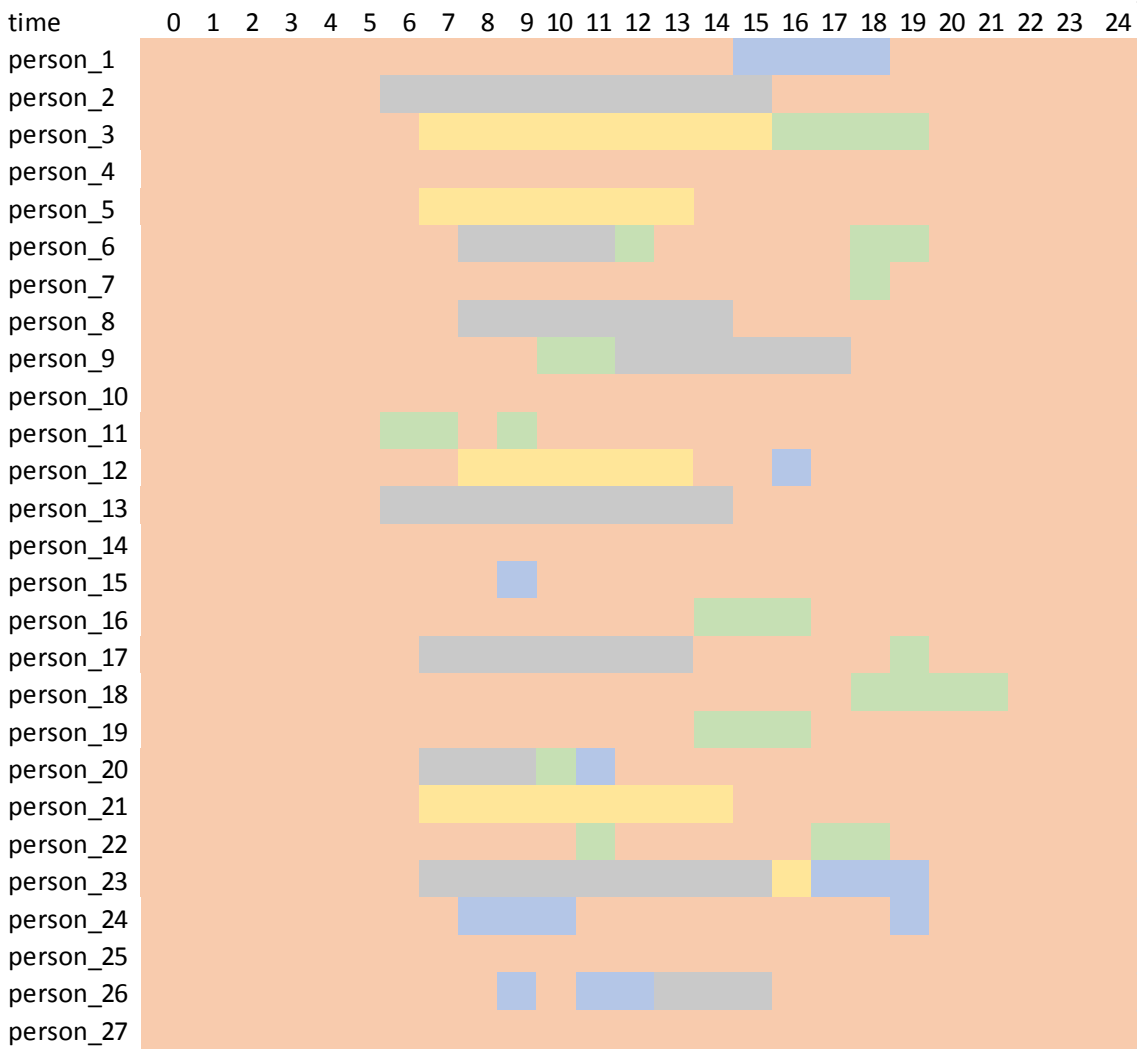


Figure 25 Snippet of the generated routines in time. Each row represents one person. Leftmost cell = 00:00, rightmost cell = 24:00. Colors: red=home, grey = work, yellow = education, green = leisure, blue = shopping/errands

Note: Due to difficulties in getting the right data for validation, the generated dataset has not been tested, however I trust the conditions that restrict the dataset according to the previous achievements.

5.5.6 Joining the Data Together

So far, we have created a set of population agents with their characteristics and home location. We also managed to generate the variety of activity patterns in accordance with the statistics from HSL and HSY for the Helsinki/Capital region. What remains is to match the agents to the generated patterns and give initial locations for the activities in order to run the model. The model would then iteratively fix the rest of the categories, namely locations of non-home activities, transport modes and transport routes.

5.5.6.1 Stratification of routines / activity patterns

To join the people to the patterns in a meaningful way, many parameters can be used as a hint. For example, children should not go to work, pensioners should only rarely have education activity and so on. Another possibility might be the number of times a person leaves home as the indication of home location, for example I suppose that people from

suburbs would tend to do less tours per day as they have fewer chances to stop by at home. However, since this has not been tested, I will only use the stratification by "stage of life" and employment status.

The following categories have been established: full-time worker, part-time worker, unemployed, little child, school-goer, student, pensioner and staying home. The last one has not been used later due to difficulties in establishing the connection with the characteristics of the agent (age, gender). Also, little children end up having no pattern as they are not usually represented in the travel survey. From the statistics for the Capital region I obtained the following distribution of the categories for Helsinki region as seen in Table 11.

Table 11 The categories and factor as a share of agent population

Category	Factor
Full time worker	0.445336976
Part time worker	0.101262469
Student (high school and older)	0.104975786
School-goer	0.079766642
Unemployed	0.069494178
Pensioner	0.198163949

Using these numbers the routines were initially assigned to these categories without any further meaning, thus for example roughly 45 % routines would end up belonging to the full-time worker category. Then, the mutation algorithm was employed to optimize to the assignment using the swap of routines between the categories. The objective function was the match to statistics for the total number of trips for each category. If the swap increased the score (how close the match is) it was kept, if not it was reversed. In this way after about 100 000 iterations a satisfactory match was found.

5.5.6.2 Stratification of agents

Even though we generated quite realistic households, the generated characteristics might not be the most useful to create the joint with the routines. Thus the agents were stratificated by the age and the location they live in, while the location would only affect the unemployment rate. The assignment function can be seen below on the Figure 26.

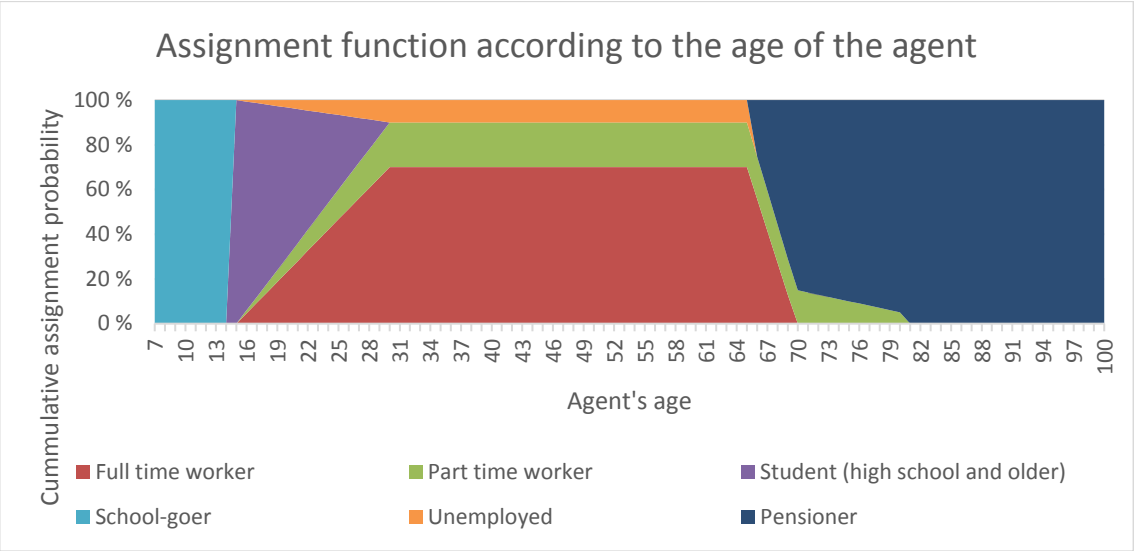


Figure 26 Assignment function according to the age of the agent, the distribution varies according to the location of the agent

5.5.6.3 Finalizing the Plans

To finalize the plans, the getter for the routine has been added to the population generator. First all the possible plans are loaded in xCity object (see Appendix 4 Population and Plans Generation Code Scheme), already stratified by our categories. Then once an agent is generated the attributes (age,location) are checked and the social category is assigned accordingly. Using this category, a random pattern from the routine patterns is picked while satisfying the category. So, for example an agent would first be assigned to student category and then he/she would pick random pattern from the student routines category. Thanks to this, it does not matter too much if the number of generated patterns is a thousand or a million, but a million might still offer slightly better variety.
Thus, agents with satisfying activity patterns are generated and the last missing attributes are the coordinates for the initial activity facility, since now we can only assign the address for home.

5.5.6.4 Facility Generation

In order to achieve realistic spatial patterns for activities I decided to use the location choice contribution of Matsim. However, this required me to generate facilities with precise locations and quite precise capacities. Using the Accessibility contribution of Matsim I was able to draw the facilities from OSM. I mostly used the assigning parameters as default, however I added medical activity to my shopping/errand activity. Unfortunately, the generator did not include any default capacities for the generated facilities, therefore I had to adjust the scripts to include them myself.

I used the following assumptions roughly based on the expected size of the facilities:

- Shop tag:
  - 500 agents for supermarket
  - 30 agents for convenience store
  - 4 agents for small shop facility
- Leisure tag:
  - 200 ice rink
  - 300 water park
  - 4 agents for small leisure facility

- Tourism tag:
    - 100 gallery
    - 200 museum
    - 800 theme park
    - 2000 zoo
    - 4 agents for small tourism facility

If building floor space was obtained, then educational facility capacity was divided by 10 and work facility by 40, otherwise both educational and work facility obtained capacity 20.
For all other cases (mostly leisure and shopping/errands) the capacity has been set to 30. Since I had the information for the available workplaces for the postcode zones data (Statistics Finland, 2015), I calibrated the workplace facility numbers accordingly. The result can be seen in Figure 27 below.

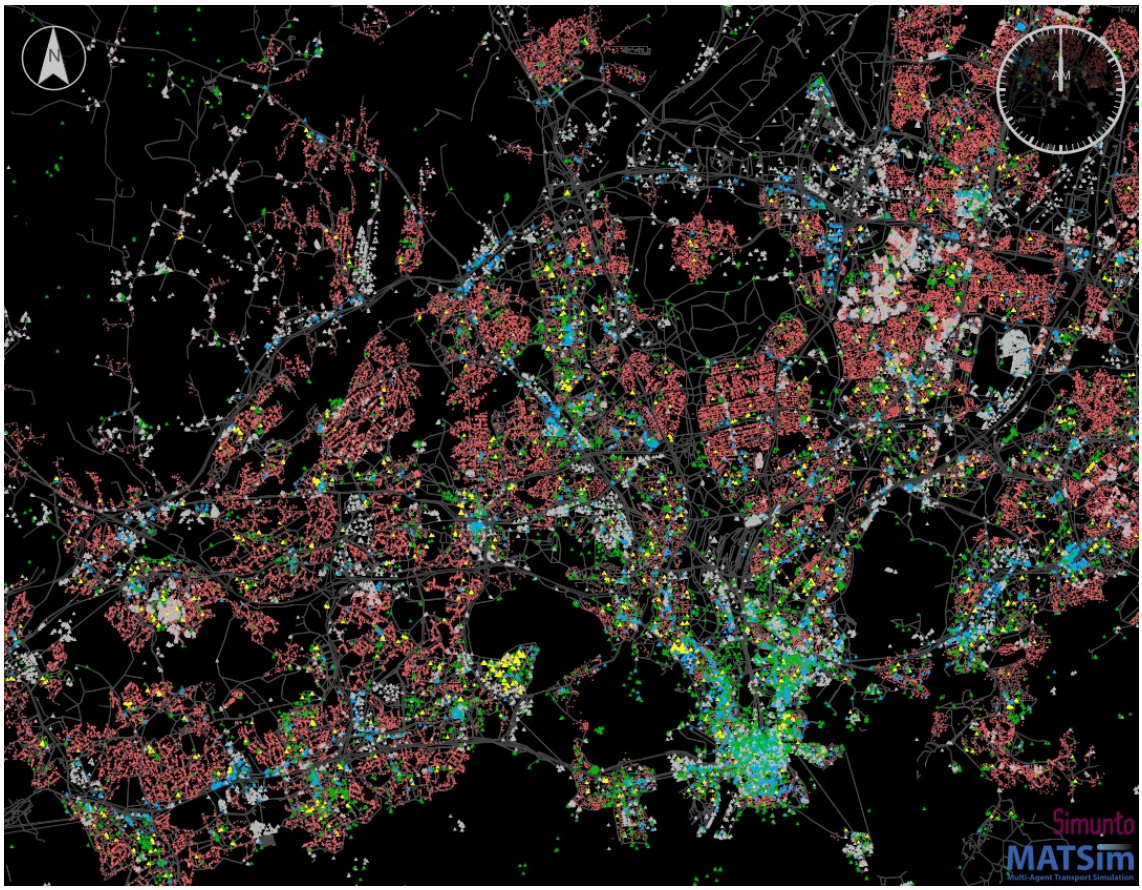


Figure 27 Facilities generated for the capital region visualized in Simunto Via. Red = housing, yellow = education, blue = shopping/errands, green = leisure, grey = work. Note that facilities might have more functions as for example most non-housing facilities also include capacity for work

The facilities generated separately are then preloaded into the population generator as well apart from the housing facilities which are created dynamically in the population generator. Thus, now the activities can use the generated facilities as a search space to assign the meaningful coordinates. Please note that educational facilities are modelled as one category which might result in some errors later. I decided for that mostly due to lack of separate information in the mobility statistics. However, this could be improved in the future.

5.5.6.5 Final scheme of the population generator

The final scheme of the population generator now includes loading of facilities and routines from separate modules, this has the advantage of better performance as well as the ability to test them separately. As a result of connecting all these modules the plans for Matsim have finally been generated. It turned out that the most efficient scheme is to always draw the full sample (the file has about 1 GB) and then draw the sample from that file if needed. As such file is quite difficult to handle is conventional ways (ordinary text editor), I do believe it can be called "synthetic big data" as it has all the desired properties of big data. The synthetic attribute has its own advantages and disadvantages which will be discussed in the end of the thesis.

5.5.7 Running the model

The logic of Matsim is to usually create a very simple initial population and then achieve the result by significant number of iterations, usually over 100, see (Väänänen, 2017) or case studies in (Horni, Nagel, & Axhausen, 2016). My approach on the other hand was to preprocess the data as much as possible and only leave the necessary adjustments to Matsim. The model managed to achieve almost the best results within about 30 iterations even for the full sample. That has some benefits, especially if running the full sample as sampling comes with its own bias in discrete simulations such as Matsim since the buses cannot really transport half of the agent, see (Horni, Nagel, & Axhausen, 2016, p. 109). And, having in mind that one the goals was to offer realistic loads for buses, it hovers the benefits of running the full sample, even if it is a computational challenge.

Another reason my approach might be beneficial is that the whole time we have the times for the trips and their categories under control, so we do not give the simulation a chance to end up with different results (unless the agents implement the plans incorrectly).

5.5.7.1 Tuning Matsim

In order to run the simulation, I needed to extend Matsim with two extra contributions. The first was SBBRaptor, a very fast router for public transport, which made the public transport computation about 100 x faster (Rieser M., 2018). The second was the locationchoice contribution which unfortunately is not part of the Matsim's core. Since some parts seemed to be a bit outdated (the implementation of public transport), I had to adjust some scripts myself to obtain the desired model.

5.5.7.2 Calibration Factors

One of the challenges is to find the right calibration factors. I decided to calibrate for modal split and average beeline trip distance for the mode as in the Table 12 below:

Table 12 The goals for calibration set from HLT 2016 (WSP Finland Oy, 2016). I calculated average trip distance by mode from the table on page 15 of HLT 2016. The values in brackets show my estimates for internal trips that seem to be more realistic, see (Laakso & Loikkanen, 2004).

Mode	Mode share	Average trip distance
Walk	27	1,5
Bicycle	6	4,5
Car	44	17 (8)
Public Transport	23	13 (9)

5.5.7.3 Matsim configuration

There are 3 variables that we need Matsim to figure out, the activity location (for others than home), the mode of transport and the route. This means that we need to estimate the right parameters for scoring in order to achieve that.

First of all, the location choice is only using 1 % sample in order to speed up simulation, this feels realistic enough given that I generated about 100 000 facilities for the Capital region. Routing approximation level is set to "noRouting" to speed up the computation again.
For the public transport, the first search radius is set to just 100 m and the extended radius is set to 300 m, again, to speed up the computation.

The biggest challenge are however the scoring parameters. I took some inspiration from (Väänänen, 2017), so I set the utilities as follows:

Table 13 (Dis-)utility table

Utility type	Walk	Cycling	Car	Public Transport
performing	+2.483			
line switch	0	0	0	-0.288
waiting	0	0	0	-4
constant	3	3,5	-2,5	+2,5
u_time	-4,5	-3,5	-2,5	-2,5

As I learned experimentally, the higher the constant compared to time disutility the more likely the mode will be used for shorter trips and vice versa. Even though I started from Väänänen's thesis I could not use the money disutility as I did not include any inputs for money in the model (no income or so), thus I am trying to use only time disutility and the constant in order to achieve simulation state close to the goals. However, perfect match to the calibration criteria was not achieved, or only partially.

5.5.7.4 Validation

The idea for validation is that the traffic volumes from the simulation should very well correlate with the observed traffic volumes. Also, the loads of the bus stops is compared to the model, as there is data available from HSL.

The comparison for the links can be seen in the following Figure 28. We can observe a correlation along the diagonal that represents perfect match, the values are quite precise on the scale of magnitude, but can be 50 % or more off from the original value. The error seems to appear equally to both sides of the diagonal, thus the source of the error might be the spatial distribution of the activities. I tried to avoid the external trips error obtained in (Zhong, Shan, Du, & Lu, 2015) by only including links within relatively central area of the model (mostly within Ring Road I), but it might still affect the model as well. Also, since the night volumes in the network are low and quite stochastic, the accuracy seems to be lower as well.

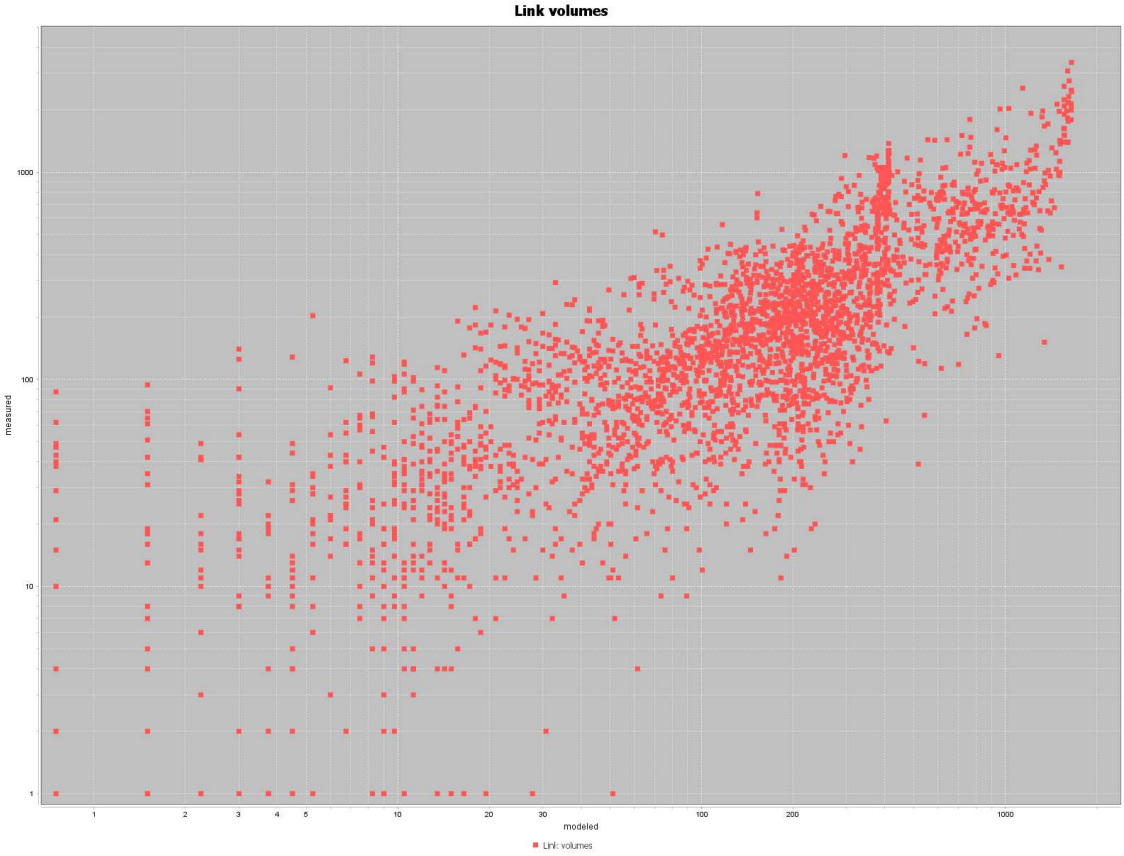


Figure 28 Hourly link volumes for Helsinki, horizontal axis modelled data vs. measured on vertical axis, logarithmic scale

For public transport stops as seen in Figure 29 the volumes tend to be lower for hubs like metro station and overestimated for the small bus stops. I suspect this is due to shorter distances of public trips than in reality. Not including commuters from outside of Helsinki might also play a role in the error of the hubs. Once again, spatial imprecision would also contribute to the distortion. Still, on the scale of magnitudes, the results are correct.

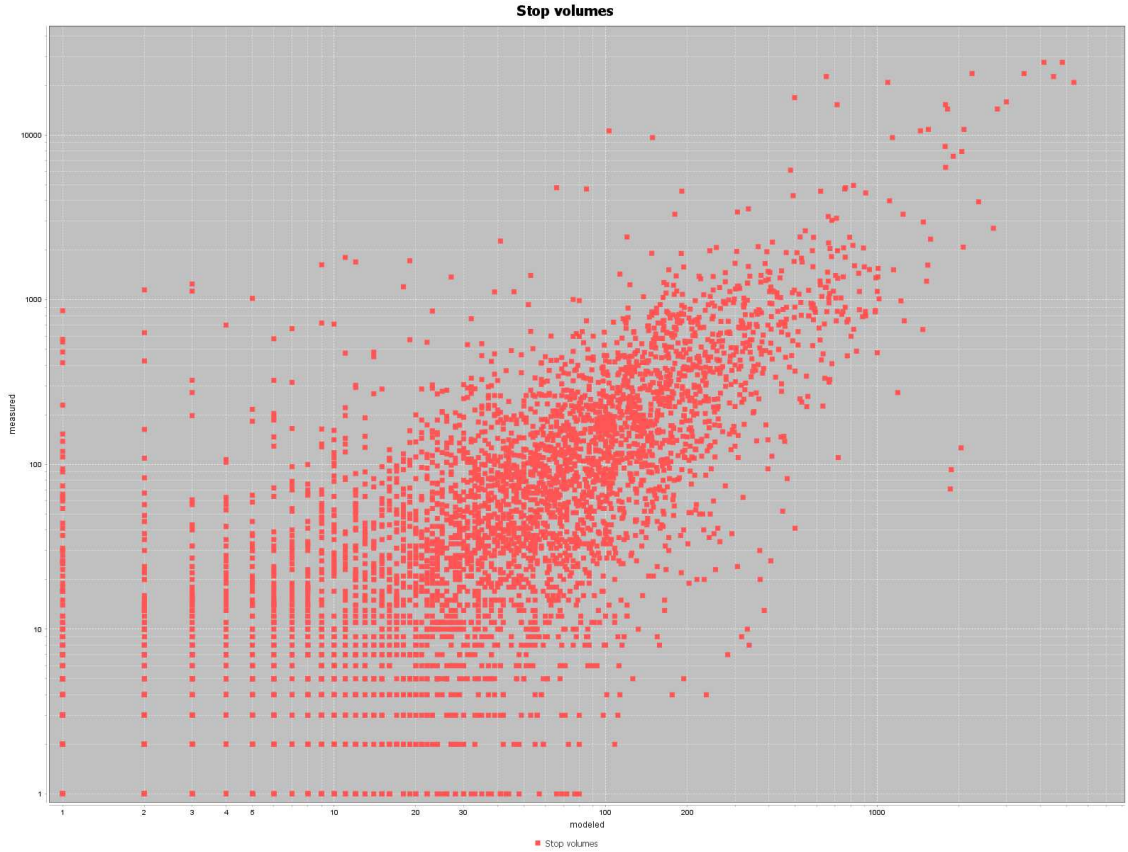


Figure 29 Daily stop passenger volumes compared to the data from HSL, model on horizontal axis, measured data on vertical. Logarithmic scale

Overall, I believe that the validation has shown the model going in the right direction, nevertheless the precision shall still be improved by following the sources of imprecisions.

6 Results and Discussion

As a result of this work, a spatio-temporal model of the Helsinki Capital region based on big data has been established using a novel approach.

6.1 Scanning the Big Data Environment

In Chapter 3 the existing big data possibilities were investigated. The most tempting big data – mobile phone data – has been used a number of times, but remains mostly inaccessible. Furthermore, when using these data the whole models tend to be restricted in use due to their higher privacy sensitivity. For the purpose of this thesis I did contact the mobile network operator Telia. The only output I would be able to get is the hourly OD matrix which would be only slight improvement to the current. Furthermore, the cost for the data would be inability to use my model commercially within VTT since one of the conditions was to use data only for the academic purposes.

As a part of the thesis other two big data sets were tested. The log of tweets from Twitter and log of Reittiopas queries. The log of tweets was found to be too sparse for Helsinki region to be useful, the queries in Reittiopas would need a complicated calibration process to become useful and it was found that it is easier to avoid using the dataset further in thesis. Additionally, it would again restrict the usage of my model, as the queries might be considered private data.

Opposite to the datasets related to the movement of people, the datasets related to the transport network such as OSM and the datasets containing public transport timetables were relatively easy to obtain and do not possess further restrictions for the model.

For the purpose of creating the model generating the synthetic population based on aggregated data appeared as the best solution. Despite hardships associated with such approach such as many detailed steps involved in the process, it allows the model to be less restricted in terms of privacy. It also turned out that such a dataset is more meaningful as we can try to establish all the possible relations between agents and the built environment. With each aggregated data measuring the population from a new angle the dataset gains more meaning and becomes more robust. This however also means having to deal with certain inconsistencies between the statistics themselves, especially when produced in different time.

6.2 Modelling the Capital Region

The model is based on three main data sources: network from Openstreetmap, public transport information obtained through GTFS and synthesized population based on the disaggregated statistics.

While I consider the first two sources to be quite reliable and complete (Barrington-Leigh & Millard-Ball, 2017) the synthetic population data naturally come with lots of errors. Generating the synthetic population for Matsim is composed of two parts, the population structure data and the movement data. The population structure data implements an object structure with the whole area acting as the main object and all the others being member of a superior object. For example, a person is a member of a family, a family is member of household which is member of a building. This approach enabled me to achieve household structures that are realistic, however the location of the households is only precise with respect to districts.

Movement data was synthetized from HSL travel survey and disaggregated with a handful of assumptions that are stated in the thesis. This method was used since I was unable to retrieve any raw data for any Finnish city, perhaps due to privacy concerns. It would still be possible to get such a data on request (as happened in case of thesis of (Väänänen, 2017)), however I aimed for developing a more flexible method that would enable me to create such model even without the proper travel survey raw data. The method I developed was to uncover the underlying tour patterns, distribute them in time while matching with other data and finally get an activity origin-destination matrix for every hour of the day. Using such a base data I constructed a graph with capacities that was then used by a set number of agents to find their daily routines in time. This yielded fairly realistic results while keeping in line with the original data. Only the activity durations were sometimes strange as there was no constraint for agents to perform a 15-hour work activity apart from the lower likelihood.

The population data was matched to the movement data using the age and employment as a key. This ensures that people of working age are far more likely to go to work instead of education, however the algorithm does not render that impossible.

Finally, a map of facilities was drawn from OSM to assign initial location of activities other than home and the capacities were assumed as no data was found to draw them precisely.
The model was set up in Matsim with several modules. To achieve more realistic spatial distribution of activities and correct travel lengths the location choice module was used. For the modal choice I modelled cars and bicycles as routed modes, public transport as partially routed and walking as teleported. All the routed modes are able to share the same links and can pass each other if the width of the link permits it.

Having a quite detailed mobility plans paid off in decreasing the need for iterations of Matsim as the scoring achieved the stabilized phase after about 20 iterations, and after 10 iterations no significant improvements from location choice were obtained so the module can be switched off.

Calibration turned out to be fairly challenging. I started from parameters used in (Väänänen, 2017) but as I did not use monetary properties of the agents I could not use these parameters directly. Only a state relatively close to the desired modal split and mode distance distributions was achieved. Achieving the optimum proved challenging especially due to the non-linearity of the model, for example a small change for walking utility would disproportionately affect all other modes and their travel distances.

Validation was made with the road volumes data and the data from the public transport stops. The data shows that the trends are right, but the spatial distributions seems to be slightly off (see Chapter 5.5.7.4 Validation), perhaps as a result of hardships with calibrating the trip distance distributions.

The purpose of the model was among others to offer a traffic data for further simulation as well as data for passenger loads. As the raw data on these are not openly available, I will show the examples from my model. In the Figure 30 we can see the passenger volumes for the 553 line going in eastern direction. The visualized traffic data can be seen in the Figure 31, with green representing low average speed and red representing high average speed.

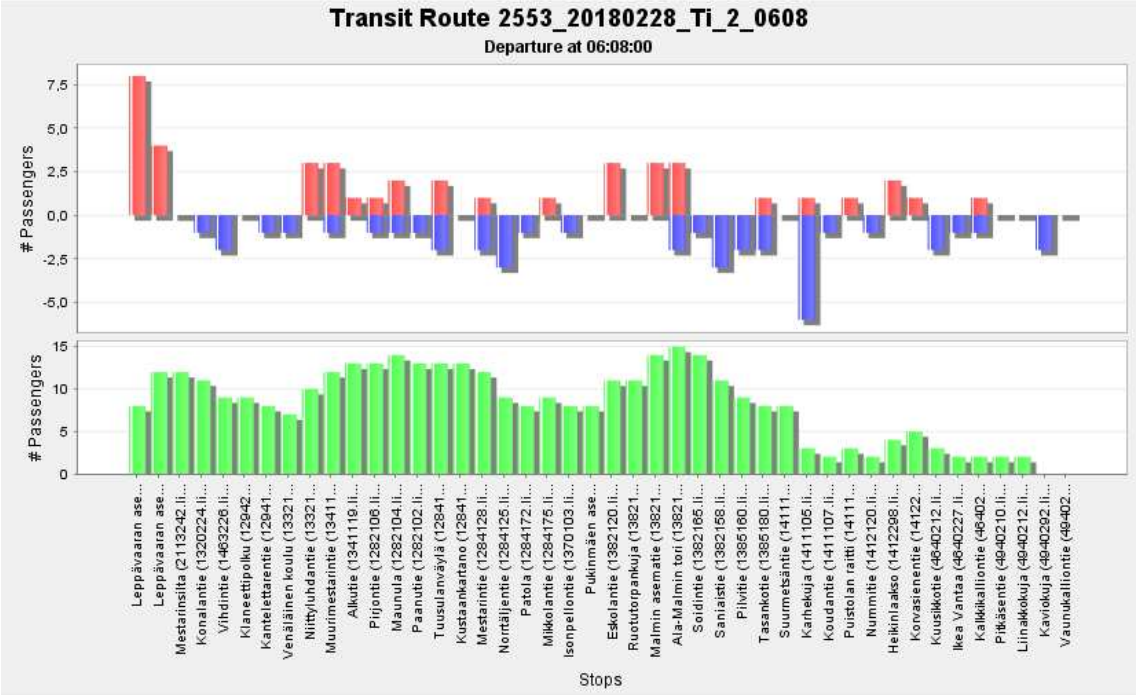


Figure 30 Passenger load of line 553, departure 6:08 from Leppävaaran asema. Upper chart shows boarding and alighting passenger, the chart below shows the load of the bus

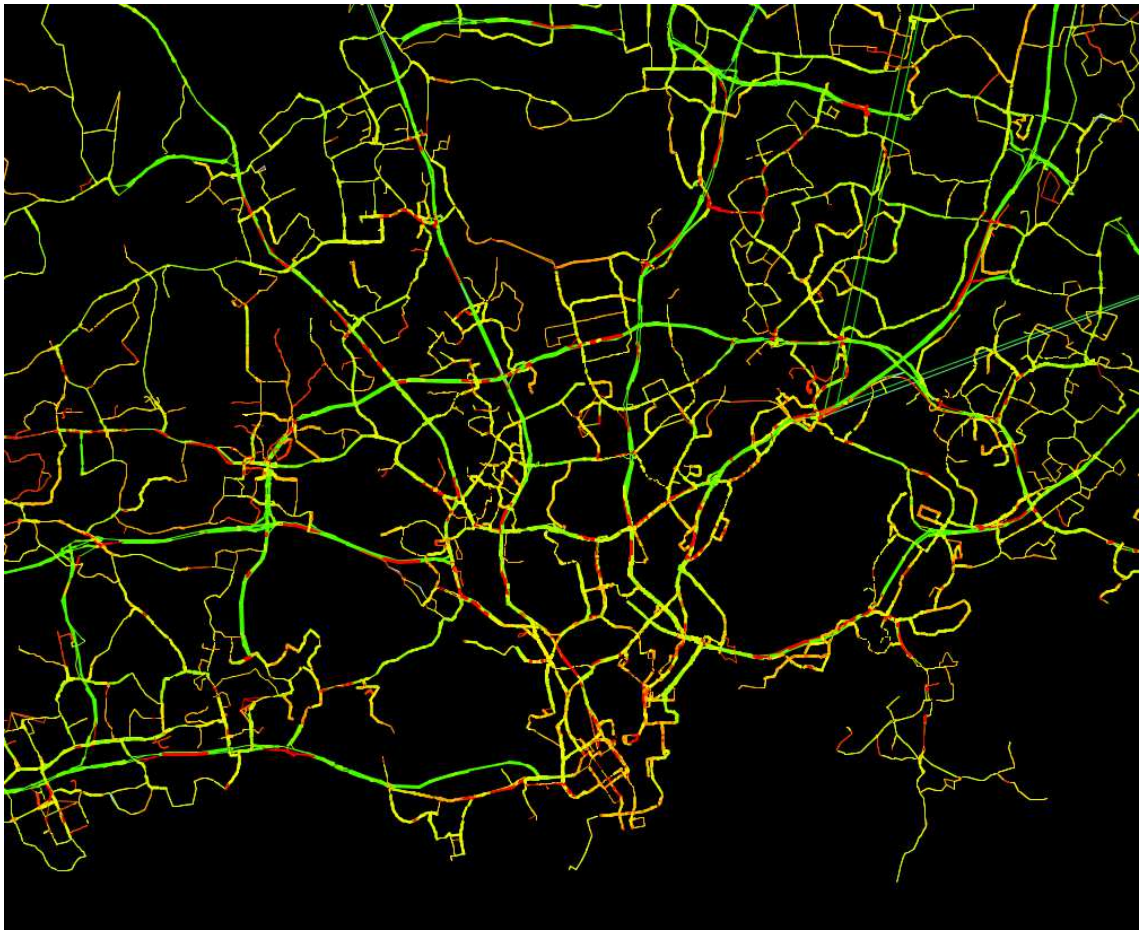


Figure 31 Average speed on the links with bus traffic at 8:02 AM, green represents high average speed, red represents low average speed (not necessarily a congestion)

6.3 Possible Improvements

The model is still limited in its scope, since it only simulates the inhabitants of Capital region moving around the network drawn from OSM framing it with a couple of kilometers of extra space. Thus, the first necessary improvement would be to include agents that arrive and leave that area during the day. Another possibility is to explicitly model tourists and other visitors. However, they are now to some extent compensated by inhabitants living in the area, but currently not being there (for example an Espoo inhabitant on a holiday in Spain) some properties are altered like the number of trips to the airport or Suomenlinna. With proper statistics this is perfectly doable.

Another missing part is the cargo traffic. Modelling such seem more challenging as I am not too familiar with its complexity. Perhaps some inspiration from existing models could be used. One thing to consider is also the scope as it would be possible to model flow of goods as well if there is a demand. This does not have a direct effect on the traffic flow, but it might have an effect on the journeys modelled for single vehicles. Furthermore, it can use the existing network of facilities to gain even more meaning.

To improve the precision of the time step model, the agents should have a preferred way to continue their "path". For that, scanning the raw data of travel survey would give a good starting point.

The location of households could also be improved by gathering better information about the buildings and iterative swapping of households within the district to achieve similarity with the population grid. This was not done in the thesis only because I was focusing on the mobility part but seems quite straightforward.

In the part matching people with mobility patterns it would be beneficial to use multiple criteria. As of now, I am only using "position in society" as a criteria composing of age and employment/student status. I believe the patterns are also influenced by location in the city, available vehicles or the position within the family. This would however require multidimensional assignment.

The model itself could be further enhanced by routing the walk legs as now it might be too easy to hop from an island to island (but one may allow that in winter if the bays are frozen). Additional modes might be tested as well, such as electric vehicles, drones etc. Matsim is quite flexible as long as the implementation for the mode behavior is provided.

Finally, there is a lot to for the performance enhancements, such as implementing faster router (hub labelling algorithm), pseudo simulation (pSim) or adapting the engine. Currently, the full sample is running one iteration in about one hour depending on the modules.

6.4 Possible Use Cases for the Developed Model

The developed model could be used to perform analysis that were unimagined before for the four-step model as it actually goes to the level of a single individual.

The obvious case is the policy impact analysis to see how the highway tolls would impact people's daily travelling and perhaps businesses. Another case might be cost-benefit analysis, since the model directly works with travel times. The monetary costs would need to be enhanced however, thus the model would need to be improved.

An innovative case might be travel experience study where we can get an information like time spent seated on the bus, time spent waiting in noisy environment or the possible delays.

Quite unexpected point might be the analysis for the business as advertising companies might be interested in how many people see the roadside advertisement and their demographic structure. Service-oriented businesses might be interested in the source of their customers and their mode split.

Once the school are properly calibrated, the city officials might be interested in the journeys to school and exposure of children to traffic. This might help them to design certain improvements where necessary for safety. The city might be as well interested in seeing the "urban experience" of an individual, for example how much time people spent around green areas and how much time they end up being frustrated in a delayed bus.

Another point might be to investigate the source of the congestion. Is it people going to work to a place with little public transport options? Is it people missing some good bus connection? Is it people living in the remote suburbs?

The model might need a little extra calibration for all these cases, but the core is ready for these investigations.

6.5 The Take-Away Message for the Big Data Environment

Another point of this thesis was to scope the possibilities for utilizing the existing big data. Openstreetmap and GTFS seem to prove very useful, but a good data for population movement data is missing. I investigated number of possibilities in the first chapter including some success stories of others, however for my case they proved not to be useful. They would always suffer from incompleteness (Twitter data, Reittiopas data), privacy issues (data from mobile phones), licensing issues (data from mobile phones) or impossibility to repeat the process because the source is not open (Reittiopas, mobile data). However, the privacy environment in Finland seems to be fairly strict, as the raw data for the travel surveys are open in other countries of the EU, such as Italy, I managed to find such data for Torino for example. I believe that it would be beneficial to many researchers and consultants to open that data to public as less assumptions would need to made and more variables could be directly measured. Still, travel surveys suffer from a certain sampling bias, especially for public transportation as discussed in chapter 5.5.7.

The closest to measuring the full sample of population is the mobile data with various subtypes and respective precisions. Opening up this data is not directly possible as the privacy of the users might be threatened. However, it would be possible to generate the data about facility loads or travelling patterns such as number of tours per day. Another question to me is if that data should be commercial by nature as it is generated by the users. The situation is quite strange in a way that the mobile companies are partially unable to legally sell the data since they never ask for permissions to do so. One of the ways to open up that market could be to explicitly ask the users for their data perhaps for a compensation. All that being said, there have been use cases for this data in activity-based models, see chapter 3.1.2.

During the model building I have realized that the joint travel modelling is very challenging without a proper survey thus it is not modelled in my model. Matsim developers made an extension Socnetsim that is supposed to model the social network. I believe that web

social networks could perhaps help to model a similar network for Matsim but I have not tested that.

Another idea is to use the real big data just to get the desired properties, like visited facilities per day. If the data we use is reliable enough we could perhaps build our model like a giant puzzle, in a similar fashion as I created the population just from the reports. The benefit of that is that the resulting model should be more robust, and less a subject to bias of a single big data source.

7 Conclusion

There were two aims for the thesis. First, to explore the possibilities of using new big data sources for activity-based models and second, to develop an activity-based model in Matsim for Helsinki. The outcome of the first task served as an initial position for the second one.

Out of the big data, several potential data sources were evaluated. Starting with mobile phone data the possibilities for usage in activity-based model were examined. As it turned out, mobile data would be beneficial source for activity-based models for its completeness and accuracy, but suffer greatly from privacy issues and the data is not open which restricts further usage of the model. For example such a model is hard to test for other researchers.

Among other, social networks seemed to offer somewhat related data to travel demand, but the data for Finland seems to be too sparse to be useful. Privacy is smaller issue as the examined data was all publicly accessible (tweets on Twitter network). Another possibility was seen it the log of queries for Reittiopas journey planner. The log of queries was however not related too well to general travel demand as people tend to search only for certain part of their journeys in the day. Another problem is the lack of other modes in the data and possible privacy issues, since the log is not part of the open data.

Other sources did not bring considerable benefits to modelling of movements of people, but sources such as Openstreetmap and GTFS turned out to be useful in building the model.

All in all, for the model it was better in the end not to use directly big data sources that were investigated in Chapter 3. Instead, I carried on with creating synthetic big data dataset by disaggregating and connecting all the available statistics into a linked dataset.

The approach can be broken down into multiple steps. First, a linked population of agents is created using the statistical data. Then, mobility patterns are generated separately by disaggregating the statistics from travel survey for Helsinki Region (HLJ, 2013). The disaggregation process yielded a Markov chain where activities by hour in the day are states and moving between these states is driven by a transition matrix varying by time. This chain can then be translated into a directed graph where the agent travels from one state to another with probability according to the transition matrix. This gives the benefit of avoiding the “last person assignment” problem. Using this graph the population for activity-based model was created.

An addition required for iterative improvement in activity locations is the knowledge for facilities. I turned to Openstreetmap to generate the facilities in the right locations and assumptions were used to assign a capacity. Furthermore, the work facilities were weighted by available postcode data from Statfin (Statistics Finland, 2015).

Having all these inputs it was possible to run Matsim model for the full sample in the Capital region of Helsinki. The model was calibrated using standard Matsim utility framework. Journeys leading outside of the region are not part of the model as well as cargo traffic. The model reached stability after about 20 iterations which can be considered an improvement to current conditions. In the end, the model was validated against the hourly link volumes for Helsinki and daily passenger volumes for stops in the Capital Region.

The validation results showed strong correlation between the modelled values and the measurements, however the values were still far from exact.

Despite all the effort, the developed model is still not perfect, therefore there are many alleys of possible future development. First, people commuting from outside of the studied area should be modelled as well and vice versa. Furthermore, tourists should have their own patterns and cargo transport should be included.

Regarding the methodology, it turned out to be useful in data environment that does not offer high quality detailed data. On the other hand, it still misses good validation, for example accounting for changes in the network. While the model itself should serve for short time-scale analysis (a few years), the methodology is more robust and could be loaded with future predictions to get an activity-based model for the future.

References

Anda, C., Erath, A., & and Fourie, P. (2017). Transport modelling in the age of big data. International Journal of Urban Sciences, 21, 19-42.

Balmer, M. (2007). Travel demand modeling for multi-agent transport simulations: Algorithms and systems. Zurich: ETH.

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. PloS one, 12(8).

Castiglione, J., Bradley, M., & Gliebe, J. (2015). Activity-based travel demand models: A primer. No. SHRP 2 Report S2-C46-RR-1.

Chaniotakis, E., Antoniou, C., & Pereira, F. (2016). Mapping social media for transportation studies. IEEE Intelligent Systems, 31(6), 64-70.

Chiba, T., Hino, H., Akaho, S., & Murata, N. (2017). Time-varying transition probability matrix estimation and its application to brand share analysis. PloS one, 12(1), e0169981. doi:https://doi.org/10.1371/journal.pone.0169981

da Silva, L., & Silva, T. (2018). Extraction and Exploration of Business Categories Signatures. Retrieved 10 26, 2018, from http://dainf.ct.utfpr.edu.br/~thiagohs/papers/daSilvabidu18.pdf

Der Standard. (2009). Mobilkom gibt Bewegungsdaten für Geo-Marketing frei. Retrieved 10 25, 2018, from https://derstandard.at/1259282147270/Mobilkom-gibt-Bewegungsdaten-fuer-Geo-Marketing-frei

Di Minin, E., Tenkanen, H., Hausmann, A., Heikinheimo, V., Järv, O., & Toivonen, T. (2016). Social media data for analysing spatio-temporal patterns and nature-based preferences of people in national parks. Helsinki: Agile 2016. Retrieved from https://agile-online.org/conference_paper/cds/agile_2016/posters/196_Paper_in_PDF.pdf

Drchal, J., Čertický, M., & Jakob, M. (2015). Data driven validation framework for multi-agent activity-based models. Istanbul: Springer.

Eunoia Project. (2012). EUNOIA project. Retrieved 10 25, 2018, from http://eunoia-project.eu/

Guido, G., Rogano, D., Vitale, A., Astarita, V., & Festa, D. (2017). Big data for public transportation: A DSS framework. 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (pp. 872-877). Napoli: IEEE.

HLJ. (2013). Liikkumistottumukset Helsingin seudulla 2012. Retrieved 10 26, 2018, from https://www.hsl.fi/sites/default/files/uploads/liikkumistottumukset_helsingin_seudulla_2012_hlj2015_raportti_0.pdf

Horn, C., Klampfl, S., Cik, M., & Reiter, T. (2014). Detecting outliers in cell phone data: correcting trajectories to improve traffic modeling. Transportation Research Record: Journal of the Transportation Research Board, 2405, 49-56.

Horni, A., Nagel, K., & Axhausen, K. e. (2016). The multi-agent transport simulation MATSim. London: Ubiquity Press. doi:10.5334/baw/

HSY. (2012, 04 02). Building (construction) information grid. Retrieved 08 08, 2018, from https://hri.fi/data/en_GB/dataset/rakennustietoruudukko

HSY. (2012). Väestöruudukko. Retrieved 08 07, 2018, from https://www.hsy.fi/sites/AvoinData/AvoinData/SYT/Tietoyhteistyoyksikko/Vaestoruudukko.pdf

Iqbal, M., Choudhury, C., Wang, P., & González, M. (2014). Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies, 40, 63-74.

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. Journal of transport geography, 52, 90-97.

Ježek, J., Jedlička, K., & Martolos, J. (2015). Visual Analytics of Traffic-Related Open Data and VGI. ICIST 2015 Conference.

Jiang, S., Ferreira, J., & González, M. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. IEEE Transactions on Big Data, 3(2), 208-219.

Joubert, J., & Van Heerden, Q. (2013). Large-scale multimodal transport modelling. Part 1: Demand generation. Retrieved 10 26, 2018, from https://researchspace.csir.co.za/dspace/bitstream/handle/10204/6963/Joubert1_2013.pdf?sequence=1&isAllowed=y

Kesting, A., Treiber, M., & Helbing, D. (2008). Agents for traffic simulation. Retrieved from https://arxiv.org/pdf/0805.0300

Kitamura, R. (1988). An evaluation of activity-based travel analysis. Transportation(15), 9-34.

Kousa, A., Matilainen, L., Koskentalo, T., Soares, J., Karppinen, A., & Kukkonen, J. (2015). Ilmansaasteille altistumisen arviointi ajankäyttötietojen avulla. In Ajassa kiinni ja irrallaan-yhteisölliset rytmit 2000-luvun Suomessa (pp. 167-182). Helsinki: Tilastokeskus.

Laakso, S., & Loikkanen, H. (2004). Liikenteen ja kommunikaation kehitys ja rakenne. In Kaupunkitalous. Helsinki: Gaudeamus.

Lappalainen, J. (2016). Journey Planner query logs as a proxy for travel demand: a case study of the Helsinki Metropolitan Area. Aalto University.

Litman, T. (2003). Measuring transportation: traffic, mobility and accessibility. ITE journal, 73(10), 28-52.

McNally, M. (2000). The four step model. Retrieved from https://cloudfront.escholarship.org/dist/prd/content/qt7j0003j0/qt7j0003j0.pdf

Mladenović MN, T. A. (2014). The Shortcomings of the Conventional Four Step Travel Demand Forecasting Process. Journal of Road and Traffic Engineering, 60(1), 5-12.

Morris, D. Z. (2015). How AT&T is using drivers' cellular data to help fix California traffic. Retrieved 10 24, 2018, from http://fortune.com/2015/10/16/att-using-big-data-to-fix-traffic/

Nurul Habib, K. E.-A. (2016). How Large is too Large? The Issue of Sample Size Requirements of Regional Household Travel Surveys, the Case of the Transportation Tomorrow Survey in the Greater Toronto and Hamilton Area. Washington DC: Transportation Research Board.

O'Fallon, C. a. (2003). Trip chaining: Understanding how New Zealanders link their travel. Institute of Transportation Engineers. ITE Journal, 73(10), 28-32.

OpenStreetMap Wiki. (2018). Automated Edits code of conduct. Retrieved 10 25, 2018, from https://wiki.openstreetmap.org/wiki/Automated_Edits_code_of_conduct

Picornell, M., Lenormand, M., Tugores, A., Dubernet, T., & Lucio, A. (2015). D6.3 Case study 3: Barcelona. EUNOIA Consortium.

Poletti, F. P. (2017). Public transit route mapping for large-scale multimodal networks. ISPRS International Journal of Geo-Information, 6(9).

Porras, P. a. (2006, 9). Large-scale collection and sanitization of network security data: risks and challenges. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.449.8616&rep=rep1&type=pdf

Positium, LBS. (2014). Feasibility study on the use of mobile positioning data for tourism statistics-consolidated report. Eurostat.

Pozdnoukhov, A. (2015, 10 30). Activity-based travel demand modeling with cellular data. Retrieved 25 10, 2018, from http://ucconnect.berkeley.edu/sites/default/files/file_uploads/pozdnukhov_caltrans_2015.pdf

Pozdnukhov, A. (2016). Demand Forecasting and Activity-based Mobility Modeling from Cell Phone Data. Retrieved from https://pdfs.semanticscholar.org/3361/ddc93a6d9f451c1b94cc488c52fd0bdf5c83.pdf

PTV Group. (2014). Ptv vissim 7 user manual.

RHYTK. (2018, 11 19). File:Helsingin seutu suomi.svg. Retrieved from Wikimedia Commons: https://commons.wikimedia.org/wiki/File:Helsingin_seutu_suomi.svg

Rieser M., M. D. (2018). Adding Realism and Efficiency to Public Transportation in MATSim. 18th Swiss Transport Research Conference. Monte Verità / Ascona.

Rinne, M. B. (2018). Automatic Recognition of Public Transport Trips from Mobile Device Sensor Data and Transport Infrastructure Information. Dublin, Ireland: Springer.

Schewel, L., & Friedman, P. (2015). USA Patent No. US20150005007A1.

Schönhof, M. a. (2007). Empirical features of congested traffic states and their implications for traffic modeling. Transportation Science, 41(2), 135-166.

Start-Up Nation Finder™. (2018). Trendit. Retrieved 10 25, 2018, from https://finder.startupnationcentral.org/company_page/trendit

Statistics Finland. (2007). Quality description: Families. Retrieved 10 26, 2018, from http://www.stat.fi/til/perh/2007/perh_2007_2008-05-30_laa_001_en.html

Statistics Finland. (2011). Time use survey. Retrieved 10 26, 2018, from http://www.stat.fi/til/akay/index_en.html

Statistics Finland. (2015). Paavo-Open data by postal code area. Helsinki.

Statistics Finland. (2018). Population census. Retrieved 10 26, 2018, from https://www.stat.fi/tup/vl2010/index_en.html

Statistics Finland. (2018). Population structure [e-publication]. Retrieved 10 20, 2018, from http://www.stat.fi/til/vaerak/meta_en.html

Statistics Finland. (2018). Taulukko- ja muuttujaluettelo. Retrieved 08 08, 2018, from https://www.stat.fi/tup/vaesto_perheet/taulukko_ja_muuttujaluettelo.html

Stopher, P. a. (2007). Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice, 41(5), 367-381.

StreetLight Data, Inc. (2018). Transportation Planning Resources. Retrieved 10 25, 2018, from https://www.streetlightdata.com/resources/#case-study

Tafidis, P. T. (2018). Can Google Maps Popular Times Be an Alternative Source of Information to Estimate Traffic-Related Impacts? Retrieved 10 25, 2018, from http://ria.ua.pt/bitstream/10773/23688/1/15.%20Tafidis%20et%20al_Can%20google%20maps%20popular%20times_TRB2018.pdf

te Brömmelstroet, M. N. (2017). Experiences with transportation models: An international survey of planning practices. Transport Policy, 58, 10-18.

Tenkanen, H. (2017). Capturing time in space: Dynamic analysis of accessibility and mobility to support spatial planning with open data and tools (1 ed.). Helsinki: Helsingin yliopisto. Retrieved from http://urn.fi/URN:ISBN:978-951-51-2935-9

Waller, M. a. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77-84.

Van Tilburg, C. (2011). Traffic policy and circulation in Roman cities. Acta Classica, 149-171.

Van Zuylen, H. a. (1980). The most likely trip matrix estimated from traffic counts. Transportation Research Part B: Methodological, 14(3), 281-293.

Ward, J. a. (2013). Undefined by data: a survey of big data definitions. Retrieved from https://arxiv.org/pdf/1309.5821

Wardrop, J. (1952). Some theoretical aspects of road traffic research., (pp. 325-378). London.

Weiner, E. (1997). Urban Transportation Planning in the United States: An Historical Overview. United States Dept. of Transportation Research and Special Programs Administration, Washington, DC.

von Mörner, M. (2017). Application of Call Detail Records-Chances and Obstacles. Transportation research procedia, 25, 2233-2241.

WSP Finland Oy. (2016). Henkilöliikennetutkimus 2016: Helsingin seutu. Retrieved from https://www.liikennevirasto.fi/tilastot/henkiloliikennetutkimus/julkaisut

Väänänen, T. (2017). An activity-based model of travel demand using an open-source simulation framework. Retrieved from https://aaltodoc.aalto.fi/bitstream/handle/123456789/29366/master_V%C3%A4%C3%A4n%C3%A4nen_Touko_2017.pdf?sequence=1

Zhong, M., Shan, R., Du, D., & Lu, C. (2015). A comparative analysis of traditional four-step. Transportation Planning & Technology, 38(5), 517-533.

Zilske, M., & Nagel, K. (2014). Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. Procedia Computer Science, 32.

Appendices

Appendix 1. Excerpt of twitter data. 1 page.
Appendix 2. Code extracting tweets with coordinates and time, plotting them afterwards. 1 page.
Appendix 3. Households in Capital Region by Size and Statistical Unit. 7 pages
Appendix 4 Population and Plans Generation Code Scheme. 1 page
Appendix 5 The distribution of work activity starts vs ends. 1 page

Appendix 1. Excerpt of twitter data

Excerpt from the tweet dump in json format

{"contributors": null, "truncated": false, "text": "I'm at \u041c\u0435\u0442\u0440\u043e \u00ab\u041c\u0435\u0436\u0434\u0443\u043d\u0430\u0440\u043e\u0434\u043d\u0430\u044f\u00bb (metro Mezhdunarodnaya) in \u0421\u0430\u043d\u043a\u0442-\u041f\u0435\u0442\u0435\u0440\u0431\u0443\u0440\u0433 https://t.co/SnC3TsFq0q", "is_quote_status": false, "in_reply_to_status_id": null, "id": 724976322997575681, "favorite_count": 0, "source": "Foursquare", "retweeted": false, "coordinates": {"type": "Point", "coordinates": [30.37968636, 59.87017926]}, "timestamp_ms": "1461682799468", "entities": {"user_mentions": [], "symbols": [], "hashtags": [], "urls": [{"url": "https://t.co/SnC3TsFq0q", "indices": [72, 95], "expanded_url": "https://www.swarmapp.com/c/ejgFQ7z2ZsP", "display_url": "swarmapp.com/c/ejgFQ7z2ZsP"}]}, "in_reply_to_screen_name": null, "id_str": "724976322997575681", "retweet_count": 0, "in_reply_to_user_id": null, "favorited": false, "user": {"follow_request_sent": null, "profile_use_background_image": false, "default_profile_image": false, "id": 1160458134, "verified": false, "profile_image_url_https": "https://pbs.twimg.com/profile_images/706751992199299072/h3B-YRiQ_normal.jpg", "profile_sidebar_fill_color": "000000", "profile_text_color": "000000", "followers_count": 214, "profile_sidebar_border_color": "000000", "id_str": "1160458134", "profile_background_color": "642D8B", "listed_count": 12, "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme10/bg.gif", "utc_offset": 10800, "statuses_count": 63316, "description": "Nastya, 18. Jeg behersker russisk (som morsm\u00e5l), engelsk og norsk #biathlon (#Russia #Italy #Norway) #Eurovision #crocheting #Coldplay #\u0401\u043b\u043a\u0430 #tennis osv.", "friends_count": 135, "location": "Saint Petersburg, Russia", "profile_link_color": "003366", "profile_image_url": "http://pbs.twimg.com/profile_images/706751992199299072/h3B-YRiQ_normal.jpg", "following": null, "geo_enabled": true, "profile_banner_url": "https://pbs.twimg.com/profile_banners/1160458134/1459007130", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme10/bg.gif", "name": "syttifemte", "lang": "ru", "profile_background_tile": true, "favourites_count": 24067, "screen_name": "n_nastya97", "notifications": null, "url": null, "created_at": "Fri Feb 08 15:27:43 +0000 2013", "contributors_enabled": false, "time_zone": "Volgograd", "protected": false, "default_profile": false, "is_translator": false}, "geo": {"type": "Point", "coordinates": [**59.87017926, 30.37968636**]}, "in_reply_to_user_id_str": null, "possibly_sensitive": false, "lang": "und", "created_at": "Tue Apr 26 14:59:59 +0000 2016", "filter_level": "low", "in_reply_to_status_id_str": null, "place": {"full_name": "\u0421\u0430\u043d\u043a\u0442-\u041f\u0435\u0442\u0435\u0440\u0431\u0443\u0440\u0433, \u0420\u043e\u0441\u0441\u0438\u044f", "url": "https://api.twitter.com/1.1/geo/id/a76c0cd7d56c4836.json", "country": "\u0420\u043e\u0441\u0441\u0438\u044f", "place_type": "city", "bounding_box": {"type": "Polygon", "coordinates": [[[30.089689, 59.776171], [30.089689, 60.092125], [30.567707, 60.092125], [30.567707, 59.776171]]]}, "country_code": "RU", "attributes":

Appendix 2. Code extracting tweets with coordinates and time, plotting them afterwards

tweets.py:

```
import json
import matplotlib.pyplot as plt,mplleaflet
import folium
fname='C:/matsim-0.9.0/Twitter data/test_finland.txt'
tweets=[]
coords=[]
colors=[]
hours=[]
map_1 = folium.Map(location=[60.22, 24.9],
                   zoom_start=12,
                   tiles='Stamen Terrain')
with open(fname, 'r') as f:
    for row in range(20000):
        if row % 100000==0:
            print(row)
        tweets.append(json.loads(f.readline()))
        try:
            #Take the coordinates from the tweet
            tc = tweets[row]['geo']['coordinates']
            #Check
            if tc[0]>60.11 and tc[0]<60.34 and tc[1]>24.48 and tc[1]<25.2:
                coords.append(tc)
                h=int(tweets[row]['created_at'].split(' ')[3].split(':')[0])
                hours.append(h)
                if h>12:
                    if h>18:
                        if h>21:
                            c='black'
                        else:
                            c='red'
                    else:
                        if h>15:
                            c='cyan'
                        else:
                            c='green'
                else:
                    if h>6:
                        c='blue'
                    else:
                        c='violet'
                if c!='violet':
                    c='black'
                colors.append(c)
                folium.Marker([tc[0],    tc[1]],    popup=tweets[row]['text'],icon=folium.Icon(color=c,icon='info-
sign')).add_to(map_1)
        except:
            pass

map_1.save(outfile='map.html')
```

Appendix 3. Households in Capital Region by Size and Statistical Unit

Legend:
Pks_kosa: my internal code for statistical zone
Pks_pop: total population
Pks_tot: total number of households
Pks_1..7: number of households of size 1..7
Source: Open data of municipalities combined together, compare with for example http://www.aluesarjat.fi/

pks_osa	location	pks_kosa	pks_pop	pks_tot	pks_1	pks_2	pks_3	pks_4	pks_5	pks_6	pks_7
111	espoo	e111	6534	3566	1741	1160	317	248	77	16	7
112	espoo	e112	7005	3589	1592	1165	409	297	97	20	9
113	espoo	e113	4268	2246	1024	732	247	192	39	8	4
114	espoo	e114	3115	1186	289	354	214	207	93	20	9
115	espoo	e115	1132	366	34	117	69	90	43	9	4
116	espoo	e116	2022	811	223	253	118	163	41	9	4
117	espoo	e117	2594	993	254	304	143	188	80	17	7
118	espoo	e118	3801	1875	793	616	202	176	68	14	6
131	espoo	e131	6869	3028	1056	929	444	424	135	28	12
132	espoo	e132	3039	1275	412	381	198	180	80	17	7
133	espoo	e133	5355	2573	1045	788	376	249	88	19	8
141	espoo	e141	3967	1514	311	535	240	309	92	19	8
142	espoo	e142	1769	913	438	258	103	76	29	6	3
143	espoo	e143	5339	1845	241	637	317	441	160	34	15
151	espoo	e151	5008	2449	1052	745	287	254	85	18	8
152	espoo	e152	3984	1669	468	575	273	249	80	17	7
161	espoo	e161	1315	605	226	189	83	81	20	4	2
211	espoo	e211	3614	2053	1024	715	157	110	36	8	3
212	espoo	e212	4027	2285	1228	660	176	170	39	8	4
213	espoo	e213	2779	1497	702	506	148	97	34	7	3
214	espoo	e214	4493	2267	957	768	247	234	47	10	4
215	espoo	e215	6362	2747	871	924	367	430	119	25	11
222	espoo	e222	3498	1885	803	744	199	98	31	7	3
231	espoo	e231	3062	1225	312	424	178	216	73	15	7
232	espoo	e232	5731	2680	1052	859	295	336	106	22	10
241	espoo	e241	4764	1765	314	620	314	364	117	25	11
242	espoo	e242	3415	1202	203	352	224	313	84	18	8
251	espoo	e251	202	68	8	24	14	10	9	2	1
252	espoo	e252	3408	1301	309	379	242	270	78	16	7
311	espoo	e311	4666	2535	1277	755	244	174	65	14	6
312	espoo	e312	5456	2614	1040	841	344	277	86	18	8
313	espoo	e313	6135	3034	1235	1065	319	297	91	19	8
314	espoo	e314	3204	1856	955	634	142	83	32	7	3
315	espoo	e315	1283	424	44	133	83	112	40	8	4
316	espoo	e316	1	1	1	0	0	0	0	0	0

321	espoo	e321	2461	1008	301	308	143	186	54	11	5
322	espoo	e322	6960	3000	903	1064	400	482	116	24	11
323	espoo	e323	6205	3359	1610	1081	350	233	65	14	6
331	espoo	e331	518	185	26	63	35	48	10	2	1
332	espoo	e332	3318	1590	617	525	213	180	42	9	4
411	espoo	e411	4372	2398	1192	759	220	155	55	12	5
412	espoo	e412	6677	3347	1407	1145	369	296	100	21	9
413	espoo	e413	7769	4171	2027	1299	421	282	109	23	10
414	espoo	e414	4019	1613	375	584	276	273	81	17	7
415	espoo	e415	1483	558	109	193	104	100	40	8	4
421	espoo	e421	1752	579	84	172	86	146	70	15	6
422	espoo	e422	2047	661	87	157	136	193	68	14	6
423	espoo	e423	3855	1911	836	590	207	197	62	13	6
431	espoo	e431	4787	1706	267	550	331	409	114	24	11
432	espoo	e432	3114	1141	188	404	218	224	82	17	8
433	espoo	e433	2514	1003	240	326	192	195	38	8	4
434	espoo	e434	4963	1786	361	495	345	403	140	29	13
441	espoo	e441	3432	1480	491	453	237	201	75	16	7
442	espoo	e442	838	330	81	112	49	62	20	4	2
443	espoo	e443	1557	569	89	201	106	130	33	7	3
451	espoo	e451	632	241	50	89	35	44	17	4	2
452	espoo	e452	4	3	2	1	0	0	0	0	0
511	espoo	e511	5052	2143	746	568	341	339	114	24	11
512	espoo	e512	578	225	50	73	48	37	13	3	1
521	espoo	e521	1896	761	222	234	108	125	55	12	5
522	espoo	e522	1982	705	174	149	129	168	65	14	6
611	espoo	e611	3650	1943	919	592	254	122	43	9	4
612	espoo	e612	2943	1243	362	408	216	189	52	11	5
613	espoo	e613	12458	5925	2515	1679	788	598	265	56	24
614	espoo	e614	1991	679	109	173	148	175	57	12	5
615	espoo	e615	1667	612	105	209	120	123	42	9	4
616	espoo	e616	2432	1093	415	305	166	144	49	10	4
621	espoo	e621	3126	1289	386	402	202	196	79	17	7
622	espoo	e622	1115	419	110	112	65	87	35	7	3
631	espoo	e631	800	331	106	94	42	70	15	3	1
632	espoo	e632	4322	1902	699	533	276	276	91	19	8
633	espoo	e633	239	90	18	32	15	16	7	1	1
634	espoo	e634	2592	978	225	296	176	187	72	15	7
635	espoo	e635	239	90	18	32	15	16	7	1	1
642	espoo	e642	936	362	75	133	61	62	24	5	2
643	espoo	e643	866	342	81	124	60	40	28	6	3
644	espoo	e644	48	23	9	8	2	3	1	0	0
645	espoo	e645	48	23	9	8	2	3	1	0	0
711	espoo	e711	373	127	22	41	18	23	17	4	2
712	espoo	e712	298	104	22	25	25	14	14	3	1

713	espoo	e713	1391	584	189	161	98	103	26	5	2
714	espoo	e714	1649	561	67	191	100	136	51	11	5
715	espoo	e715	3394	1203	246	333	216	253	119	25	11
721	espoo	e721	2534	994	256	313	175	153	74	16	7
722	espoo	e722	196	107	54	33	9	7	3	1	0
723	espoo	e723	972	334	38	122	62	69	33	7	3
724	espoo	e724	210	135	83	36	10	5	1	0	0
725	espoo	e725	196	107	54	33	9	7	3	1	0
10	helsinki	h010	7395	4015	1904	1344	401	271	68	13	14
20	helsinki	h020	639	365	166	146	37	13	1	1	1
30	helsinki	h030	1036	553	243	201	63	34	7	4	1
40	helsinki	h040	11938	7234	4008	2289	535	304	68	19	11
50	helsinki	h050	9130	5568	3162	1677	419	220	71	11	8
60	helsinki	h060	1094	565	266	166	67	39	23	4	0
70	helsinki	h070	10784	6096	3078	2004	502	398	93	12	9
80	helsinki	h080	4426	2262	911	868	244	170	52	12	5
90	helsinki	h090	473	221	71	91	30	18	9	1	1
101	helsinki	h101	7439	4685	2703	1467	325	139	37	12	2
102	helsinki	h102	2356	1212	480	469	144	93	22	4	0
103	helsinki	h103	0	0	0	0	0	0	0	0	0
104	helsinki	h104	0	0	0	0	0	0	0	0	0
111	helsinki	h111	2471	1527	813	552	105	47	9	1	0
112	helsinki	h112	9437	6489	4187	1841	316	110	31	3	1
113	helsinki	h113	6982	5249	3824	1180	193	43	7	2	0
121	helsinki	h121	7397	5495	3947	1277	202	58	9	1	1
122	helsinki	h122	4484	3243	2265	784	135	51	7	0	1
130	helsinki	h130	14465	8028	3892	2750	740	447	150	28	21
140	helsinki	h140	15210	9195	5135	2839	662	430	93	26	10
150	helsinki	h150	5099	3138	1878	841	216	146	40	12	5
161	helsinki	h161	2765	1602	877	452	140	107	21	4	1
162	helsinki	h162	7174	3604	1631	1100	432	271	98	31	41
171	helsinki	h171	4976	2650	1139	998	302	149	41	13	8
172	helsinki	h172	0	0	0	0	0	0	0	0	0
173	helsinki	h173	3762	2116	1088	668	199	103	32	13	13
174	helsinki	h174	125	68	27	29	9	2	1	0	0
180	helsinki	h180	1986	1226	704	363	96	49	13	0	1
190	helsinki	h190	23	10	2	5	1	2	0	0	0
201	helsinki	h201	2974	1444	569	487	208	123	38	8	11
202	helsinki	h202	28	24	21	2	1	0	0	0	0
203	helsinki	h203	7039	3534	1409	1242	488	307	77	8	3
204	helsinki	h204	1248	751	394	256	66	31	4	0	0
211	helsinki	h211	6041	3479	1837	1074	315	189	36	21	7
212	helsinki	h212	0	0	0	0	0	0	0	0	0
213	helsinki	h213	306	250	196	52	2	0	0	0	0
220	helsinki	h220	9686	5938	3302	1871	493	221	32	14	5

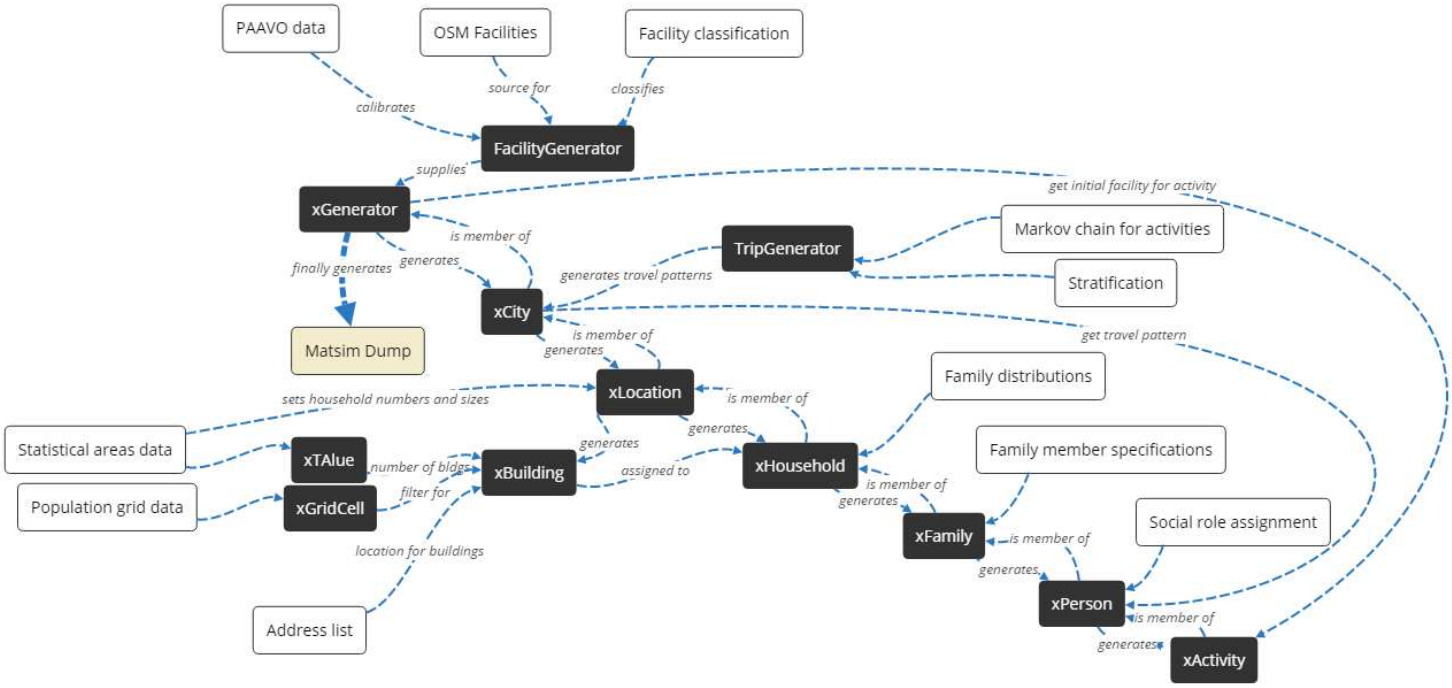
231	helsinki	h231	1483	818	424	240	77	46	24	5	2
232	helsinki	h232	7486	3653	1394	1312	465	380	67	27	8
240	helsinki	h240	3848	2016	944	630	234	138	40	20	10
250	helsinki	h250	7767	4189	2042	1267	468	307	79	18	8
260	helsinki	h260	3143	1664	841	442	211	106	38	11	15
270	helsinki	h270	783	401	187	119	39	41	13	2	0
281	helsinki	h281	486	203	41	92	30	31	7	2	0
282	helsinki	h282	7316	4286	2474	1104	343	258	79	18	10
283	helsinki	h283	1088	468	148	151	80	58	21	9	1
284	helsinki	h284	10090	5352	2497	1747	556	381	134	22	15
285	helsinki	h285	2808	1294	462	444	169	161	45	9	4
286	helsinki	h286	0	0	0	0	0	0	0	0	0
287	helsinki	h287	1389	660	286	181	94	63	18	9	9
291	helsinki	h291	12103	7562	4667	1869	550	359	94	19	4
292	helsinki	h292	870	542	327	139	42	31	3	0	0
293	helsinki	h293	9467	5485	3079	1484	483	295	97	23	24
294	helsinki	h294	4380	2517	1345	731	250	143	40	5	3
301	helsinki	h301	8490	4624	2342	1351	438	366	99	23	5
302	helsinki	h302	555	201	35	70	29	48	15	2	2
303	helsinki	h303	1123	514	184	176	63	64	21	5	1
304	helsinki	h304	5026	2920	1646	771	243	205	43	10	2
305	helsinki	h305	1275	752	454	170	61	43	20	2	2
306	helsinki	h306	1143	569	246	179	66	54	19	5	0
311	helsinki	h311	8086	4639	2475	1365	410	306	73	8	2
312	helsinki	h312	8041	4196	1805	1496	440	365	78	10	2
313	helsinki	h313	7064	3762	1819	1144	372	312	98	16	1
314	helsinki	h314	0	0	0	0	0	0	0	0	0
320	helsinki	h320	6200	3232	1515	962	389	265	79	15	7
331	helsinki	h331	13228	7346	3830	2143	694	465	143	42	29
332	helsinki	h332	2655	1035	252	340	163	195	63	15	7
333	helsinki	h333	8597	4159	1583	1487	583	322	126	33	25
334	helsinki	h334	2749	1095	188	474	203	168	44	14	4
335	helsinki	h335	271	118	30	49	20	12	7	0	0
336	helsinki	h336	94	56	35	10	7	2	2	0	0
341	helsinki	h341	6955	2770	654	990	436	489	160	30	11
342	helsinki	h342	3324	1301	286	469	212	243	65	15	11
351	helsinki	h351	5963	2220	350	825	399	499	117	25	5
352	helsinki	h352	2713	1051	204	391	192	188	60	13	3
353	helsinki	h353	288	92	9	24	24	23	7	3	2
354	helsinki	h354	10	7	4	3	0	0	0	0	0
361	helsinki	h361	1280	631	230	243	80	69	7	1	1
362	helsinki	h362	9552	4364	1725	1243	597	559	161	44	35
363	helsinki	h363	1140	626	285	220	78	34	9	0	0
364	helsinki	h364	2670	1274	524	380	179	130	41	16	4
370	helsinki	h370	8280	4511	2190	1445	465	299	77	21	14

381	helsinki	h381	6446	3647	1984	1008	340	213	57	26	19
382	helsinki	h382	6068	3078	1443	885	351	248	110	27	14
383	helsinki	h383	7224	3973	1985	1223	409	252	79	12	13
384	helsinki	h384	6	5	4	1	0	0	0	0	0
385	helsinki	h385	2292	1025	356	338	154	113	43	16	5
386	helsinki	h386	2677	1496	763	461	152	81	28	5	6
391	helsinki	h391	8103	3508	992	1329	533	476	130	36	12
392	helsinki	h392	6095	2982	1260	911	384	311	90	15	11
401	helsinki	h401	7702	3480	1135	1255	513	417	117	36	7
402	helsinki	h402	8804	4188	1640	1382	566	391	140	45	24
403	helsinki	h403	3370	1390	346	518	230	210	65	14	7
411	helsinki	h411	6711	2706	631	1011	452	427	135	31	19
412	helsinki	h412	2951	1106	198	410	199	208	59	15	17
413	helsinki	h413	16	10	6	3	0	1	0	0	0
414	helsinki	h414	5263	2689	1259	807	301	201	63	38	20
415	helsinki	h415	1462	667	246	205	108	72	24	10	2
420	helsinki	h420	3803	1871	836	541	216	186	68	15	9
431	helsinki	h431	8353	4902	2739	1355	454	264	65	14	11
432	helsinki	h432	7500	4299	2369	1140	433	266	66	17	8
433	helsinki	h433	712	408	195	150	42	16	3	2	0
434	helsinki	h434	9093	4677	2053	1577	544	342	108	25	28
440	helsinki	h440	2205	912	240	335	127	155	42	7	6
451	helsinki	h451	6061	2687	867	954	377	339	113	25	12
452	helsinki	h452	4917	2825	1559	799	245	135	55	14	18
453	helsinki	h453	4268	2419	1315	670	241	116	47	19	11
454	helsinki	h454	12047	5874	2540	1761	730	560	182	62	39
455	helsinki	h455	1954	765	154	294	135	127	37	12	6
456	helsinki	h456	18	10	7	2	0	0	0	0	1
457	helsinki	h457	4663	2490	1270	708	265	141	48	28	30
461	helsinki	h461	1868	1150	678	307	101	53	7	2	2
462	helsinki	h462	1064	532	223	183	60	45	13	6	2
463	helsinki	h463	4999	2527	1145	775	274	226	75	21	11
464	helsinki	h464	342	141	44	46	16	22	9	3	1
465	helsinki	h465	3439	1911	948	631	171	114	28	13	6
471	helsinki	h471	13816	7279	3599	2141	738	483	179	79	60
472	helsinki	h472	7479	3478	1319	1120	493	361	130	38	17
473	helsinki	h473	8564	4610	2256	1432	472	303	90	33	24
474	helsinki	h474	5153	2342	927	683	331	236	96	40	29
475	helsinki	h475	2715	1320	562	405	178	107	39	17	12
480	helsinki	h480	16	8	3	3	1	1	0	0	0
491	helsinki	h491	12090	6303	2914	1991	664	536	147	34	17
492	helsinki	h492	3084	1165	225	427	189	221	76	15	12
493	helsinki	h493	622	341	157	117	44	19	2	1	1
494	helsinki	h494	378	138	23	47	28	25	13	2	0
495	helsinki	h495	1631	697	208	239	106	104	32	5	3

500	helsinki	h500	1	1	1	0	0	0	0	0	0
510	helsinki	h510	395	159	46	50	19	30	13	0	1
520	helsinki	h520	741	311	94	97	57	43	12	6	2
531	helsinki	h531	1	1	1	0	0	0	0	0	0
532	helsinki	h532	1	1	1	0	0	0	0	0	0
533	helsinki	h533	0	0	0	0	0	0	0	0	0
541	helsinki	h541	13896	7429	3672	2177	802	532	169	48	29
542	helsinki	h542	0	0	0	0	0	0	0	0	0
543	helsinki	h543	7	5	3	2	0	0	0	0	0
544	helsinki	h544	5286	2453	951	793	326	230	91	38	24
545	helsinki	h545	6820	3318	1466	978	416	260	117	42	39
546	helsinki	h546	7689	4011	1723	1436	436	317	84	7	8
547	helsinki	h547	4297	1923	690	613	278	221	85	14	22
548	helsinki	h548	0	0	0	0	0	0	0	0	0
549	helsinki	h549	0	0	0	0	0	0	0	0	0
550	helsinki	h550	525	201	49	72	29	24	19	2	6
560	helsinki	h560	28	14	6	5	0	3	0	0	0
570	helsinki	h570	62	26	10	8	2	3	0	3	0
580	helsinki	h580	431	132	10	36	27	38	14	3	4
591	helsinki	h591	794	229	15	43	43	85	36	5	2
592	helsinki	h592	186	63	6	24	14	10	5	3	1
1	kauniainen	k01	2289	1055	408	343	112	122	54	11	5
2	kauniainen	k02	2766	1146	345	389	136	175	78	16	7
3	kauniainen	k03	1751	701	195	229	108	93	59	12	5
4	kauniainen	k04	1999	762	191	249	95	137	69	15	6
10	vantaa	v10	779	286	37	112	60	55	16	4	2
11	vantaa	v11	1630	696	193	253	112	103	29	4	2
12	vantaa	v12	8059	4142	1858	1334	438	381	101	20	10
13	vantaa	v13	3953	1851	698	620	233	207	76	11	6
14	vantaa	v14	2399	1104	348	442	147	130	21	11	5
15	vantaa	v15	16395	9072	4623	2821	798	541	194	63	32
16	vantaa	v16	5025	2788	1425	828	288	167	71	6	3
17	vantaa	v17	11701	5887	2511	1990	655	498	167	44	22
18	vantaa	v18	2768	1190	306	471	206	152	41	9	5
20	vantaa	v20	1831	697	136	252	125	129	36	13	6
21	vantaa	v21	800	445	211	159	44	20	8	2	1
22	vantaa	v22	903	528	284	155	51	34	4	0	0
23	vantaa	v23	5868	2531	816	826	347	391	121	20	10
24	vantaa	v24	1211	427	70	138	75	101	27	11	5
25	vantaa	v25	270	106	23	41	17	15	7	2	1
26	vantaa	v26	288	110	25	35	18	22	9	1	0
30	vantaa	v30	341	120	14	48	21	21	13	2	1
31	vantaa	v31	294	115	30	34	22	20	5	3	1
32	vantaa	v32	688	276	64	104	50	39	8	7	4
33	vantaa	v33	876	358	114	98	59	59	18	7	3

34	vantaa	v34	408	173	49	63	29	19	9	3	1
40	vantaa	v40	4763	1625	231	462	341	430	116	30	15
41	vantaa	v41	0	0	0	0	0	0	0	0	0
50	vantaa	v50	3472	1617	590	526	238	212	41	7	3
51	vantaa	v51	10229	4598	1678	1359	713	618	176	36	18
52	vantaa	v52	464	256	122	89	27	12	2	3	1
53	vantaa	v53	0	0	0	0	0	0	0	0	0
60	vantaa	v60	5001	2622	1201	845	294	202	65	10	5
61	vantaa	v61	5839	3423	1809	1140	258	140	49	18	9
62	vantaa	v62	5779	3087	1411	1030	378	195	51	15	7
63	vantaa	v63	6380	3253	1375	1143	366	268	68	22	11
64	vantaa	v64	2851	1148	282	404	202	174	64	15	7
65	vantaa	v65	8169	4105	1823	1239	514	371	119	26	13
66	vantaa	v66	1363	617	203	208	115	64	21	4	2
67	vantaa	v67	4129	1719	405	678	289	259	69	13	6
68	vantaa	v68	2213	942	224	397	155	120	31	10	5
69	vantaa	v69	149	53	8	19	9	12	3	1	1
70	vantaa	v70	4071	2042	858	692	231	190	55	11	5
71	vantaa	v71	4789	1806	389	598	292	367	115	30	15
72	vantaa	v72	4047	1958	871	528	267	193	60	26	13
73	vantaa	v73	2902	1108	230	406	171	206	59	24	12
74	vantaa	v74	8141	4540	2429	1274	409	275	99	36	18
75	vantaa	v75	3984	1465	263	508	269	286	95	29	15
80	vantaa	v80	2289	866	129	369	155	137	54	15	7
81	vantaa	v81	7335	3560	1457	1135	473	346	104	30	15
82	vantaa	v82	3004	1582	771	460	183	105	41	15	7
83	vantaa	v83	6245	3066	1323	910	407	301	86	26	13
84	vantaa	v84	2514	1006	247	355	164	167	49	16	8
85	vantaa	v85	1340	449	59	130	90	122	31	11	6
86	vantaa	v86	3724	1264	172	375	241	349	91	24	12
87	vantaa	v87	1432	573	133	226	83	80	34	11	6
88	vantaa	v88	1693	701	221	203	121	100	39	11	6
90	vantaa	v90	0	0	0	0	0	0	0	0	0
91	vantaa	v91	5616	2888	1352	859	338	226	66	31	16
92	vantaa	v92	0	0	0	0	0	0	0	0	0
93	vantaa	v93	2930	1308	427	461	188	163	54	10	5
94	vantaa	v94	11215	5527	2482	1602	665	491	186	67	34
95	vantaa	v95	4022	1684	396	662	310	235	61	13	7
96	vantaa	v96	2730	1078	223	409	194	188	38	17	9
97	vantaa	v97	2078	762	137	252	155	149	46	15	8
98	vantaa	v98	646	232	38	80	43	48	14	6	3

Appendix 4 Population and Plans Generation Code Scheme



Commentary:
The label with black background denote separate class, out of these only xGenerator, FacilityGenerator and TripGenerator are not instantiated.
All the labels with white background represent input, be it CSV files or just a list of numbers.
The only yellow label represents the only output of the whole process, population dump for Matsim.
Note: the dumping is simplified in this scheme, as it does not shows the xStats class

Appendix 5 The distribution of work activity starts vs ends

HWH		Start hour																							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Finish hour	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	1	0	2	7	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	13	0	0	0	1	10	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	14	0	0	0	1	2	25	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	15	0	0	0	1	2	10	80	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	16	0	0	0	0	2	22	20	75	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	0	0	0	0	10	15	15	5	15	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	18	0	0	0	0	0	12	5	3	3	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0
	19	0	0	0	0	0	0	2	0	0	1	5	1	0	0	0	0	0	0	0	0	0	0	0	0
	20	0	0	0	0	0	0	0	0	0	0	5	2	2	0	0	0	0	0	0	0	0	0	0	0
	21	0	0	0	0	0	0	0	0	0	3	0	1	3	0	0	0	0	0	0	0	0	0	0	0
	22	0	0	0	0	0	0	0	0	0	4	0	1	0	4	0	0	0	0	0	0	0	0	0	0
	23	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4	0	0	0	0	0	0	0	0	0
	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total		0	0	0	4	26	94	149	92	29	14	15	6	5	4	4	0	0	0	0	0	0	0	0	0